DOCUMENT RESUME

ED 141 864 CS 501 770

TITLE Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, January 1-March 31, 1977.

INSTITUTION Haskins Labs., New Haven, Conn.
REPORT NO SR-49 (1977)
PUB DATE 77
NOTE 248p.

EDRS PRICE MF-$0.83 HC-$12.71 Plus Postage.
DESCRIPTORS *Acoustic Phonetics; *Articulation (Speech); *Artificial Speech; *Audition (Physiology); Auditory Perception; *Beginning Reading; *Language Research; Lateral Dominance; Phonemes; Phonetics; Psychoacoustics

ABSTRACT

This report is one of a regular series about the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The 11 papers discuss the dissociation of spectral and temporal cues to the voicing distinction in initial stopped consonants; perceptual integration and selective attention in speech perception; phonetic recoding and reading difficulty in beginning readers; interactive experiments with a digital pattern playback; the function of strap muscles in speech; the genichyoid and the role of the strap muscles; syllable synthesis; articulatory movements in vowel-consonant-vowel sequences; measuring laterality effects in dichotic listening; a simple model of response selection in the dichotic two-response paradigm; and acoustic correlates of perceived prominence in unknown utterances. Other sections list publications and reports from Haskins Laboratories and the Defense Documentation Center and the ERIC numbers of earlier reports. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED141864

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SR-49 (1977)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 January - 31 March 1977

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports).

CS 501 770

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development,
Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

United States Army, Department of Defense
Contract MDA 904-77-C-0157

National Institutes of Child Health and Human Development
Contract NO1-HD-1-2420

National Institutes of Health
General Research Support Grant RR-5596

National Science Foundation
Grant BNS76-82023

National Science Foundation
Grant MCS76-81034

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer
Peter Bailey[1]
Fredericka Bell-Berti*
Gloria J. Borden*
Robert Crowder*
James E. Cutting*
Donna Erickson
Carol A. Fowler*
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Alice Healy*
Isabelle Y. Liberman*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[2]
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Donald P. Shankweiler*
Linda Shockey
George N. Sholes
Michael Studdert-Kennedy*
Quentin Summerfield[1]
Michael T. Turvey*
Robert Verbrugge*

Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Harriet Greisser*
Donald Hailey
Terry Halwes
Elly Knight*
Sabina D. Koroluk
Agnes McKeon
Nancy R. O'Brien
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Leonard Szubowicz
Edward R. Wiley
David Zeichner

Students*

Steve Braddon
David Dechovitz
Laurel Dent
Susan Lea Donald
F. William Fischer
Hollis Fitch
Anne Fowler
Nieba Jones
Lynn Kerr
Morey J. Kitzman
Andrea J. Levitt
Roland Mandler
Leonard Mark
Nancy McGarr
Georgia Nigro
Mary Jo Osberger
Sandra Prindle
Abigail Reilly
Robert Remez
Helen Simon
Emily Tobey
Betty Tuller
Harold Tzeutschler
Michele Werfelman

---

*Part-time
[1]Visiting from The Queen's University of Belfast, Northern Ireland.
[2]Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

On the Dissociation of Spectral and Temporal Cues to the Voicing Distinction in Initial Stop Consonants -- Quentin Summerfield and Mark Haggard. 1

Perceptual Integration and Selective Attention in Speech Perception: Further Experiments on Intervocalic Stop Consonants -- Bruno H. Repp. . 37

Phonetic Recoding and Reading Difficulty in Beginning Readers -- Leonard S. Mark, Donald Shankweiler, Isabelle Y. Liberman, and Carol A. Fowler 71

Interactive Experiments with a Digital Pattern Playback -- Patrick W. Nye, Franklin S. Cooper, and Paul Mermelstein 87

The Function of Strap Muscles in Speech: Pitch Lowering or Jaw Opening? -- James E. Atkinson and Donna Erickson 97

The Geniohyoid and the Role of the Strap Muscles -- Donna Erickson, Mark Liberman and Seiji Niimi 103

Syllable Synthesis -- Ignatius G. Mattingly 111

Articulatory Movements in VCV Sequences -- Thomas Gay 121

Measuring Laterality Effects in Dichotic Listening -- Bruno H. Repp . . 149

A Simple Model of Response Selection in the Dichotic Two-Response Paradigm -- Bruno H. Repp 187

Acoustic Correlates of Perceived Prominence in Unknown Utterances -- Jane H. Gaitenby and Paul Mermelstein 201

II. Publications and Reports

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-48)

On the Dissociation of Spectral and Temporal Cues to the Voicing Distinction in Initial Stop Consonants*

Quentin Summerfield and Mark Haggard†

ABSTRACT

It has been claimed that a rising first-formant (F_1) transition is an important cue to the voiced-voiceless distinction for syllable-initial, prestressed stop consonants in English. Lisker (1975) has pointed out that the acoustic manipulations suggesting a role for F_1 have involved covariation of the onset frequency of F_1 with the duration, and hence the frequency extent of the F_1 transition; he has also argued that effects hitherto ascribed to the transition are more properly attributed to its onset. Two experiments are reported in which F_1 onset frequency and F_1 transition duration/extent were manipulated independently. The results confirm Lisker's suggestion that the major effect of F_1 in initial voicing contrasts is determined by its perceived frequency at the onset of voicing and show that a periodically excited F_1 transition is not, per se, a positive cue to voicing. In a third experiment, the frequencies at the onset of voicing of both F_1 and

---

*A partial summary of these results was presented at the 90th meeting of the Acoustical Society of America, San Diego, California, November 1975. This paper has been accepted for publication in the Journal of the Acoustical Society of America.

†The Medical Research Council Hearing Research Institute, Nottingham, England.

Acknowledgment: Experiment I was conducted in the Department of Psychology at the Queen's University of Belfast, Northern Ireland with the support of grant AT/2058/021/HQ from the Joint Speech Research Unit, U.K. and grant B/RG/1466 from the Science Research Council, U.K. It was reported as "First formant onset frequency as a cue to the voicing distinction in prestressed, syllable-initial stop-consonants," in Speech Perception No. 5, pp. 25-33. (Progress Report, Department of Psychology, The Queen's University of Belfast). This paper was written, and the later experiments were carried out, at the Haskins Laboratories, New Haven, Connecticut, U.S.A. while Quentin Summerfield was supported by a N.A.T.O. postdoctoral research fellowship. We should like to express our appreciation to Alvin Liberman for his generous hospitality and encouragement, and to Bruno Repp, Peter Bailey, Gary Kuhn and David Pisoni for their criticisms of earlier drafts of this manuscript.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

F_2 were manipulated. The influence on the perception of stop-consonant voicing that resulted was determined specifically by the frequency of F_1, rather than by the overall distribution of energy in the spectrum. The results demonstrate a complementary relationship between perceptual cue sensitivity and production constraints: in production, the VOT characterizing a particular stop-consonant varies inversely with the degree of vocal tract constriction, and hence the frequency of F_1 required by the phoneme following the stop; in perception, the lower the frequency of F_1 at the onset of voicing, the longer the VOT that is required to cue voicelessness. In this way, the inclusion of F_1 onset frequency in the cue-repertoire for voicing reduces the noninvariance problem for perception.

INTRODUCTION

Lisker and Abramson (1964) suggested that the articulatory basis for the voiced-voiceless distinction for stop-consonants resides in the relative timing of laryngeal and supralaryngeal articulations. Prestressed, syllable-initial voiced stops in English display temporal coincidence of oral release with the onset of laryngeal vibration. When the onset of vocal cord vibration follows oral release by more than about 40 msec, the stop is voiceless. By translating variation on this articulatory dimension into variation of the parametric input to an acoustic speech synthesizer, Lisker and Abramson (1967) generated VOT[1] continua that spanned the two perceptual categories of voicing for each of the three places of stop production used in English. Phoneme boundaries on these continua occurred close to those values of VOT that optimally segregate voiced from voiceless stops in the productions of English speakers. Since then, VOT continua have been used extensively as experimental devices. They permit the determination of a phoneme boundary, changes in whose position can be used as sensitive indices of the perceptual consequences of variation of parameters both intrinsic (for example, Stevens and Klatt, 1974) and extrinsic (for example, Eimas and Corbit, 1973; Summerfield, 1975a) to the test syllables themselves. However, it has not always been clear which aspects of the stimulus are held to be perceptual cues, given that many of the acoustical parameters so far asserted to possess cue value have tended to covary. Incorporating covariation in a set of stimuli is well justified from an articulatory point of view if the objectives of an experiment are linguistic or cognitive. But, if the objectives are psychoacoustical or perceptual, then the use of covarying parameters begs the question of what acoustical variables are registered and

---

[1]With reference to the acoustics of production, the term 'VOT' will refer to the time interval between the onset of the occlusion release transient and the onset of quasiperiodicity. With reference to continua of synthetic stimuli, the term 'VOT' will refer to the interval between the onset of the stimulus (that may or may not include a burst) and the onset of periodic excitation. During this interval, the presence of noise excitation in F_2, F_3 and the higher formants, and the absence of energy in F_1 is implied. The term 'separation interval' will refer only to the temporal aspect of VOT.

contribute to the perception of the contrast. A precise specification of the perceptually pertinent parameters is important if valid interpretations are to be made of data obtained using various types of continua whose members are said to vary in "VOT".

Using synthetic stimuli, Summerfield and Haggard (1974) artificially varied the temporal separation of the fricated burst from the events that normally follow it: formant transitions and the onset of periodicity. They demonstrated that the temporal interval is indeed a powerful perceptual cue, whether or not it is filled with aspiration. The question remains: which of the spectral parameters of VOT whose variation is normally correlated with that of the separation interval are also perceptual cues? Stevens and Klatt (1974) suggested that some threshold duration or spectral extent of first formant (F_1) transition may be psychoacoustically a more basic cue to the voiced value of the feature, and that VOT (that is, the temporal separation interval) is grafted onto this through learning in infancy. Summerfield and Haggard (1974) showed that the detectability of transitions in both the first and higher formants, whether or not they were periodically excited, could provide important secondary cues for adults. Lisker (1975) has argued that the simple articulatory basis of VOT (for example, Lisker and Abramson, 1971) renders it the most general and basic cue, but proposed that if any secondary aspect of the acoustical array related to formant transitions is important, then it is the onset frequency of F_1 rather than its dynamic spectral properties. Lisker's data show that when the importance of the spectral cues is assessed by trading them against VOT, which in turn affects the values of the secondary transition cues, then VOT does emerge as the most potent perceptual cue. However, his results, based on a nonorthogonally varying stimulus set, implicate the average frequency region of F_2 as a functioning secondary cue in addition to F_1 onset frequency. The experiments reported here were designed to refine and extend Lisker's conclusion and to reduce the ambiguity by using orthogonally varying stimulus arrays. The matter can be simplified by asking three questions. Does F_1 onset cue a voiced percept in inverse relation to its frequency? Is a rising F_1 transition a positive cue to voicing independent of its onset frequency? Are spectral influences on the perception of voicing a function only of the frequency of F_1 or of the distribution of energy in both F_1 and the higher formants? Experiments I and II were designed to answer the first two of these questions. Experiment III was designed to answer the third question.

EXPERIMENT I: Conditions 1 and 2.

In the first condition of Experiment I, the frequency of a fixed-frequency, transitionless F_1 was systematically lowered across a set of consonant-vowel (CV) VOT continua. If Lisker's (1975) conclusion is correct, this procedure should increase the probability of a voiced percept at any given VOT. In the second condition, the onset frequency of F_1 was held constant independently of the realized VOT, while the duration, and consequently the spectral extent of F_1 transition following voicing onset, were systematically increased. If a periodically excited F_1 transition is, per se, a cue to voicing, then this procedure should increase the probability of a voiced percept at any given VOT.

Stimuli and Procedure

Both conditions of Experiment I were run interactively with stimuli generated at run-time by a Fonema OVE-IIIb serial resonance speech synthesizer controlled by the SPEX program (Draper, 1973) running on a D.E.C. PDP-12 digital computer. Stimuli were exemplars drawn from /g-k/ VOT continua spanning the VOT range from 0 msec to +80 msec in 1-msec steps. The closed-loop algorithm controlling stimulus presentation was an implementation of PEST (Taylor and Creelman, 1967) with the following control parameters: deviation limit of the sequential test (W)=0.5[2]; starting step size = 16 msec; terminating step size = 1 msec. These parameters result in an estimate of the p 0.5 point on the psychometric function underlying the physical test continuum; this point corresponds to the phoneme boundary. To achieve a controlled estimate of the position of the boundary, two PEST runs were randomly interleaved with starting points randomly drawn from preselected ranges approximately evenly balanced on either side of the subject's expected phoneme boundary region. The two interleaved runs converged independently from starting points at long and short VOTs, and subjects were unaware of performing in a closed-loop situation. Convergence was continued until the step size of each run had diminished to less than or equal to 1 msec and the VOTs corresponding to the p 0.5 estimates from each run were within 5 msec of one another. The phoneme boundary position is here defined as the average of these two independent estimates. Previously, open-loop and closed-loop procedures for estimating phoneme boundaries have been compared and shown to produce highly similar results (Summerfield, 1974a).

The stimuli used in each condition were constructed from seven five-formant CV 'stimulus types'. A stimulus type is that set of synthesis control parameters that generates a stimulus with a VOT of 0 msec. The frequency contours of F_2 and F_3 did not differ between stimulus types and were constructed with initial formant transitions appropriate for the velar place of articulation. These transitions were linear in frequency/time over their duration of 44 msec. The F_2 transition had its onset at 2400 Hz and reached a steady state at 2000 Hz. The F_3 transition had its onset at 2600 Hz and reached a steady state at 3000 Hz. F_4 and F_5 were set to 3500 Hz and 5000 Hz, respectively. The total duration of each stimulus type was 320 msec. The seven stimulus types used in Condition 1 were distinguished by the frequencies of their first formants that were set to 200, 225, 250, 275, 300, 350, and 400 Hz. The seven stimulus types used in Condition 2 were distinguished by the duration of their F_1 transitions; these transitions always onset at 250 Hz and rose linearly at 5 Hz per msec for either 0, 6, 12, 18, 24, 30 or 36 msec after voicing onset. No other synthesis control parameters were varied between stimulus types or conditions. Over the first 80 msec of each stimulus type, the overall amplitude contour was constant and the fundamental frequency (F_0) was fixed at 100 Hz so that differences in F_0 at voicing onset could not accompany differences in VOT. A stimulus with any VOT in the range 0 msec to +80 msec could be constructed from any one of the

[2]With W=0.5 the PEST algorithm is simplified. The Wald sequential decision test is obviated and a change in stimulus value occurs after every response.

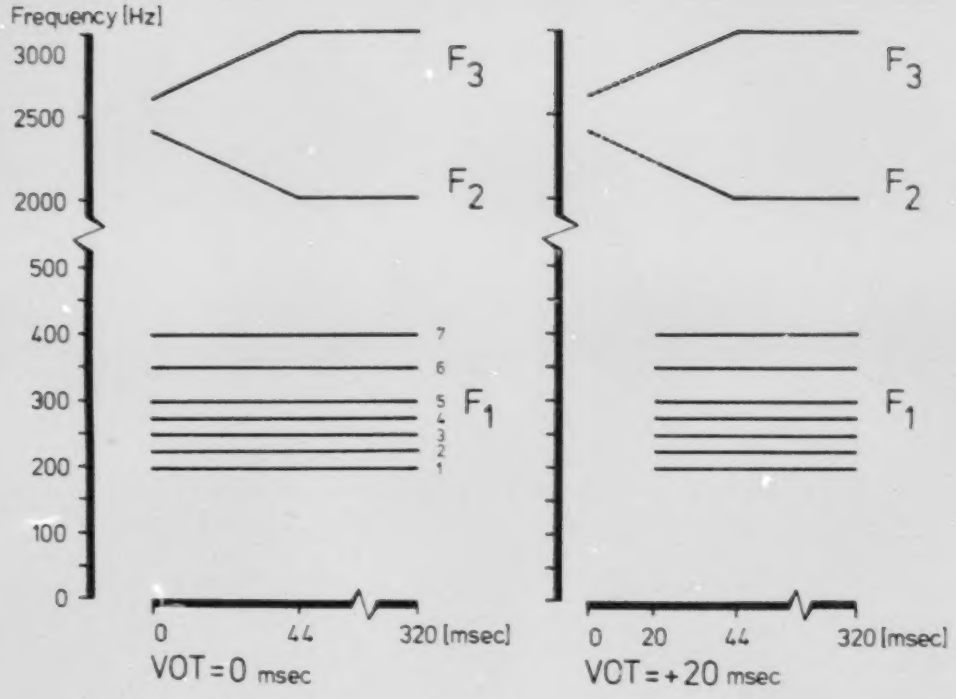


Figure 1: Schematic spectrograms showing the patterns of the first three formants for the seven stimulus types used in Experiment I, Condition 1 in exemplars with VOTs of 0 msec (left) and +20 msec (right). Solid lines indicate periodic and dotted lines aperiodic formant excitation. The stimulus types are distinguished by the frequencies of their transitionless first formants.

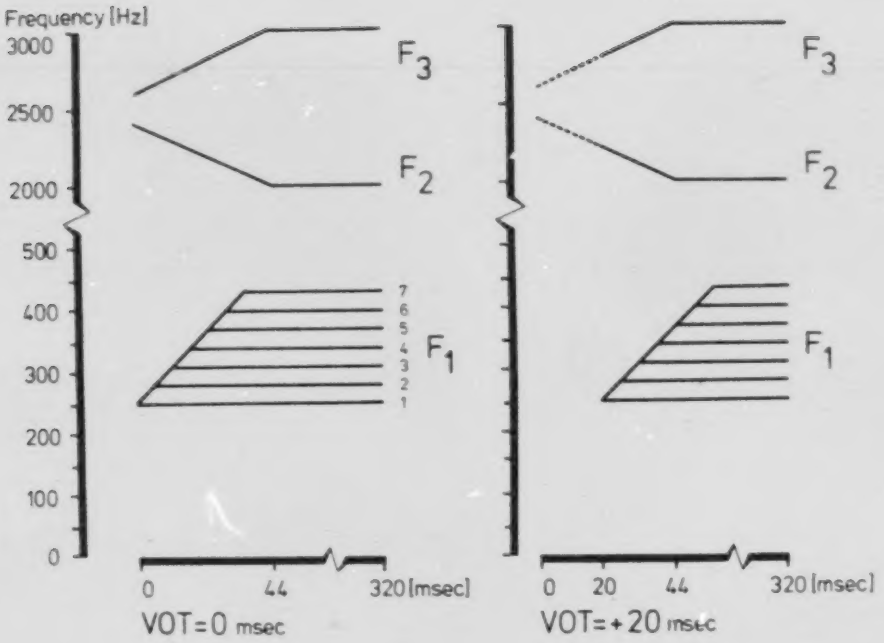


Figure 2: As Figure 1 for the seven stimulus types used in Experiment I, Condition 2. The stimulus types are distinguished by the duration and extent of their first formant transitions which onset, independently of VOT, at 250 Hz.

stimulus types by algorithm. The algorithm replaced the periodic excitation prior to the specified VOT with noise excitation 5.5 dB lower and also widened the bandwidth of F_1 from 60 Hz to 300 Hz for this portion of the syllable. This procedure reduces the level of aperiodic energy in F_1 and thereby simulates the acoustic consequences of coupling the pharynx and the trachea. The onset of pitch-pulsing was synchronized to the specified VOT by the procedure described by Draper and Haggard (1974). Figure 1 illustrates the differences between the stimulus types used in Condition 1 in displays of the formant parameter specifications F_1, F_2 and F_3 of exemplars with VOTs of 0 msec and +20 msec. Figure 2 displays analogous patterns for the stimuli used in Condition 2 and shows that in order to hold the onset frequency of F_1 constant as VOT varied, it was necessary to restructure the spectral relation between F_1 and the higher formants in a manner that is not representative of any naturally occurring variation.

Six adult subjects performed in the experiment, three in the order Condition 1 - Condition 2, and three in the reverse order. Each was a native speaker of British English and had served previously in experiments involving closed-loop phoneme boundary estimation. Stimuli were presented binaurally through AKG K60 600-Ohm headphones to subjects who sat in a sound-damped cubicle. The peak intensity of presentation was constant across subjects at approximately 85 dB SPL for stimuli with 0 msec. VOT derived from the two identical stimulus types (Types 3 and 1 in Conditions 1 and 2, respectively). Subjects were instructed to identify the initial consonant of each stimulus as either /g/ or /k/ and to indicate their response by pressing one of two buttons labeled 'G' and 'K'. A third button, labeled '?', could be pressed to summon a repetition of the current stimulus. Each subject ran through the whole set of continua twice. In Condition 1, three subjects experienced the continua in ascending, followed by descending, order of F_1 frequencies, and three in descending, followed by ascending order. The two estimates obtained were averaged to provide a single estimate for each subject on each continuum. Analogous order balancing was employed in Condition 2. The lack of naturalness inherent in the stimulus structure posed no difficulty for listeners, although some subjects reported hearing stimuli with long VOTs and extensive F_1 transitions in Condition 2, as initiated by the cluster /kl/, rather than by the single consonant /k/.

Results

The seven boundary positions obtained for each subject in each condition are plotted against the frequency of F_1 for Condition 1 in Figure 3, and against both the duration of the F_1 transition and the frequency of the F_1 steady state for Condition 2 in Figure 4. Mean boundary positions obtained by averaging these data over subjects are tabulated in Table 1 for Condition 1 and in Table 2 for Condition 2.

The results of Condition 1 support Lisker's (1975) conclusion that the onset frequency of F_1 can function as a voicing cue: the data in Table 1 show that the position of the phoneme boundary averaged across subjects decreases monotonically as the frequency of a transitionless first formant is raised. Only subject 6 failed to show an overall decrement. The seven phoneme boundaries from each of the six subjects were examined together in a

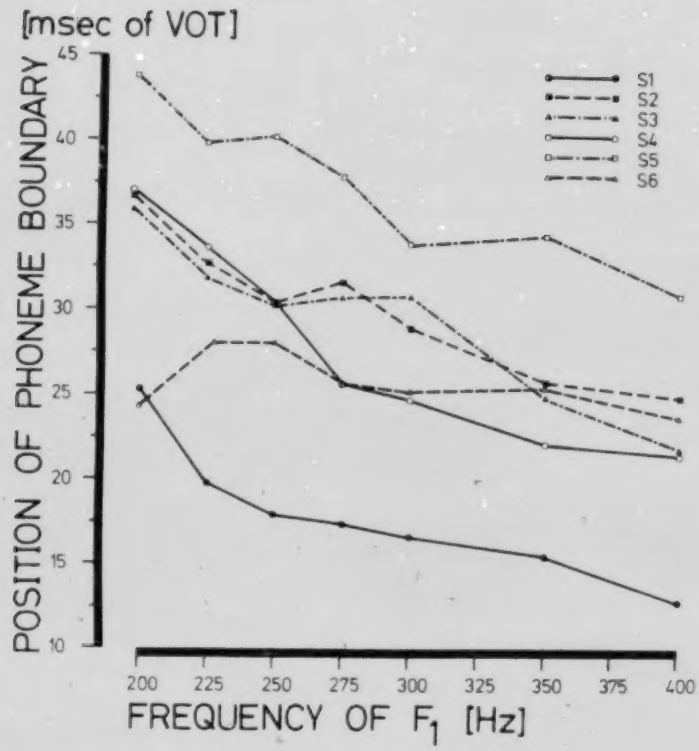


Figure 3: Results of Experiment I, Condition 1 for six individual subjects. Each point plots the mean of four phoneme boundary estimates (derived from two pairs of interleaved closed-loop estimates). Points corresponding to each subject have been connected by straight lines, showing that for five of the subjects, the voicing boundary shifted to shorter VOTs as the onset frequency of F_1 increased.

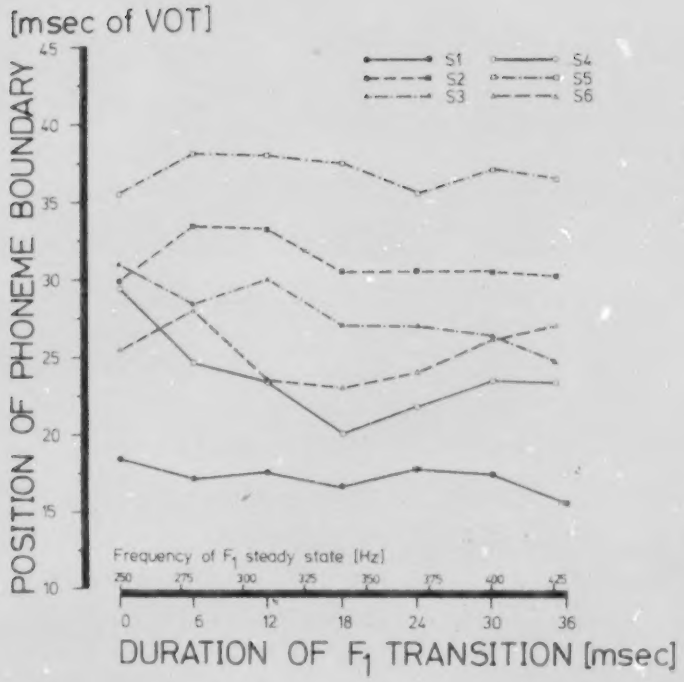


Figure 4: Results of Experiment I, Condition 2 plotted as for Figure 3. The functions for individual subjects are either horizontal (S2, S5) or decline (S1, S3, S4, S6) as the duration of the F_1 transition increased, showing that the presence of an F_1 transition does not predispose voiced percepts when its onset frequency is fixed.

TABLE 1: Experiment 1: Condition 1.

Mean phoneme boundaries in msec of VOT /PBs/ averaged over two estimates by each of 6 subjects on seven /g-k/ VOT continua differentiated by the frequency of a constant frequency, transitionless first formant (200 Hz - 400 Hz).

	Number and first formant frequency (Hz) of Stimulus Type :-						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	200	225	250	275	300	350	400
[PBs]	33.81	30.99	29.53	28.13	26.64	24.73	22.59

TABLE 2: Experiment I: Condition 2

Mean phoneme boundaries in msec of VOT /PBs/ averaged over two estimates by each of 6 subjects on seven /g-k/ VOT continua differentiated by the durations of their first formant transitions (0 msec-36 msec) that onset at 250 Hz independently of VOT.

	Number and first formant transition duration (msec) of Stimulus Type :-						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	0	6	12	18	24	30	36
[PBs]	28.25	28.01	27.59	25.74	26.04	26.81	26.18

nonparametric test for monotonic trend (Ferguson, 1966) that gave a value of the normal deviate equal to 6.19, indicating that the trend is significant ($p < 0.01$; 2-tailed). The results of Condition 2 indicate that variation in F_1 transition duration/extent does produce a small effect on the perception of stop voicing. However, it is not in the direction predicted from the arguments of Stevens and Klatt (1974) or Summerfield and Haggard (1974) on the basis of transition detectability. Table 2 shows that a fall in the value of VOT at the phoneme boundary occurred as transition duration increased. This trend is evident in the data of Subjects 1, 3, 4 and 6 and is also significant ($z=2.58$; $p < 0.05$; 2-tailed).

Discussion

The results of Experiment I imply that the critical aspect of F_1 for the perception of stop-voicing is its perceived frequency at the onset of voicing, and suggest that an F_1 transition as such does not specifically predispose a voiced percept. However, the relative amplitudes in the outputs of a serial resonance synthesizer are not fixed, but vary according to the formant frequency separations (c.f. Fant, 1960). In natural productions, constricting the supralaryngeal vocal tract lowers the frequency of F_1 and reduces the amplitudes of the higher formants and the overall intensity of the output. Increasing the frequency of F_1 in an OVE synthesizer raises the overall intensity of the output, including the higher formants, so that the distribution of energy in the spectrum increasingly favours higher frequencies. Accordingly, the results of Experiment I could reflect perceptual sensitivity either to changes in the location of the first spectral peak at the onset of periodicity, or alternatively, to changes in the amplitude of that peak relative to peaks at higher frequencies. To determine which interpretation is more appropriate, a control experiment was run using stimuli generated on a parallel formant synthesizer whose formant amplitudes could be specified individually and for which, therefore, the frequency of F_1 and the relative amplitudes of the first three formants could be varied independently.

EXPERIMENT II: Control Conditions 1 and 2

In the first control condition, nine VOT continua were created by combining each of three values of F_1 onset frequency with each of three extents of F_1 transition. Within each continuum, the onset frequency of F_1 was held constant as in Condition 2 of Experiment I. If the results of that condition reflect perceptual sensitivity to changes in the onset frequency of F_1, then phoneme boundaries should vary here with F_1 onset frequency, but not with F_1 transition extent. In the second control condition, the amplitude of F_1 relative to F_2 was varied over a 12 dB range across three VOT continua, while the spectral specification of the stimuli comprising the continua was unchanged. If the results of Experiment I reflect perceptual sensitivity to changes in relative formant amplitudes, then phoneme boundaries should shift to shorter VOTs as the intensity of F_1 is reduced relative to F_2. Alternatively, if the results reflect sensitivity to the frequencies of spectral peaks at the onset of periodicity, rather than to their absolute or relative amplitudes, then the three boundaries should coincide.

Control Condition 1: Stimuli and Procedure

Nine two-formant /g-k/ VOT continua were synthesized on the parallel resonance synthesizer at Haskins Laboratories (Mattingly, 1968). Each continuum consisted of eight 300 msec stimuli that varied in VOT from +15 msec to +50 msec in 5 msec steps with the onset of pitch pulsing synchronized to the intended VOT. As VOT increased along each continuum, the amplitude of F_1 was reduced to zero and F_2 was excited by noise. Stimuli with the same VOT in different continua were differentiated only by the frequencies of their first formants. Within any continuum the actual onset frequency of F_1 was fixed and did not vary with VOT. Nine continua were created by combining three values of F_1 onset frequency (208, 311 and 412 Hz) with three frequency extents of F_1 transition (200, 100 and 0 Hz). The duration of these transitions was 20 msec. (The first formant frequency parameter changed over five successive 5-msec intervals, reaching a steady state in the fifth interval.) The transition rates were, therefore, 10 Hz/msec, 5 Hz/msec and 0 Hz/msec. The transition rate of 5 Hz/msec is the same as that used in Condition 2 of Experiment I. The transition duration of 20 msec is longer than the 15 msec that Stevens and Klatt (1974) showed to be the 75 percent threshold duration for detection of an F_1 transition changing at a rate of 8.5 Hz/msec. The acoustic differences among the members of the continua are exemplified in Figure 5, where the formant parameter specifications of stimuli in which F_1 onsets at 208 Hz with VOTs of 0 msec and +20 msec are displayed.

Two groups of subjects listened to a randomization that included ten occurrences of each of the 72 stimuli. Stimuli were presented binaurally through Grason-Stadler TDH39-300Z headphones at a level of 85 dB SPL (peak deflection). One group of subjects consisted of six members of the research staff of Haskins Laboratories, any of whose residual phonetic naivety was dispelled by a description of the acoustic structure of the stimuli. The other group consisted of nine students attending a Yale University summer school who declared themselves to be phonetically naive. Subjects were instructed to make a forced choice identification of the initial consonant of each stimulus as either /g/ or /k/, but to indicate in addition if the sound that they heard was not a satisfactory exemplar of a CV syllable initiated by either /g/ or /k/.

Control Condition 1: Results

Four of the experienced subjects and six of the naive subjects exhibited predictable performance: they reported few instances of stimuli initiated by phonemes other than /g/ or /k/ and reported increasing numbers of /k/ percepts as VOT increased along each continuum. However, the VOT range +15 msec to +50 msec was not sufficient to permit the computation of a phoneme boundary for every subject in every condition. Accordingly, the data from Condition 1 are summarized in Table 3 not as phoneme boundary positions, but as percentages of /g/ responses made by these ten subjects to the eight members of each continuum combined. Figure 6 displays plots of the percentage of /g/ responses made to each stimulus in each continuum averaged across these subjects. Each point plots the mean of 100 observations.

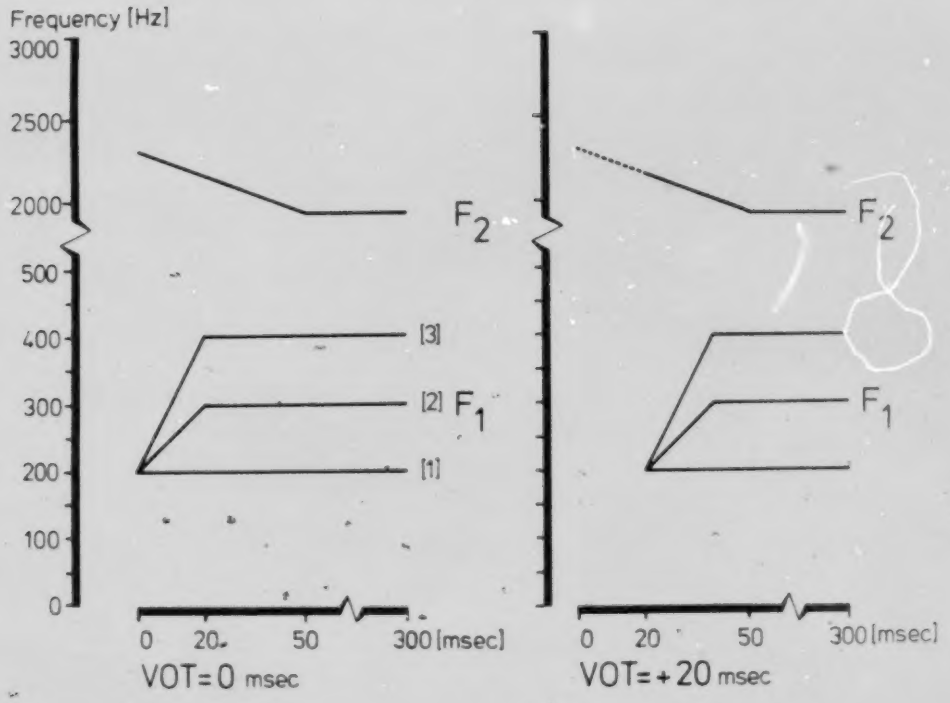


Figure 5: Schematic spectrograms showing the pattern of the first two formants for stimuli used in Experiment II, Condition 1 in exemplars with VOTs of 0 msec (left) and +20 msec (right) in which F_1 onsets at 208 Hz. Stimuli were derived from nine VOT continua distinguished by a) the onset frequencies of their first formants (208, 311 or 412 Hz), and b) the extent of their first formant transitions (0, 100 or 200 Hz).

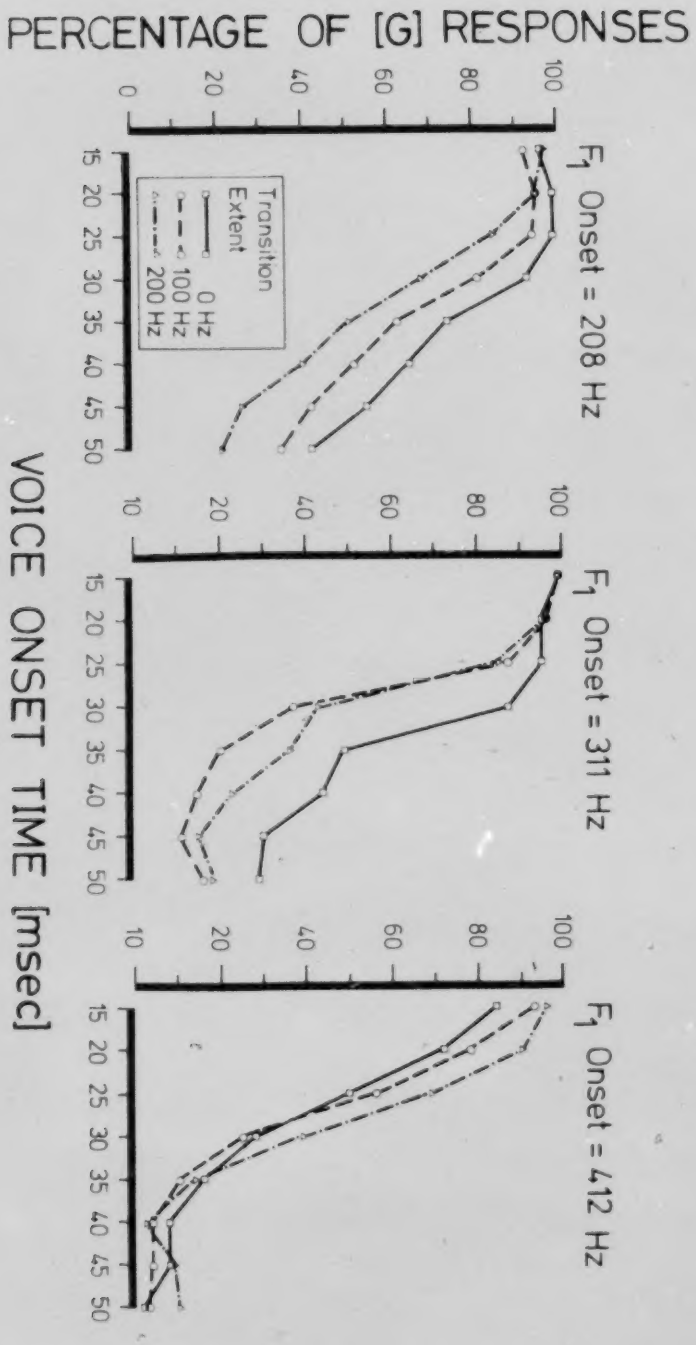


Figure 6: Results of Experiment II, Condition 1 pooled over 10 subjects.

14

TABLE 3 : Experiment II : Condition 1.

Percentages of 'G' responses made to the members of nine /g-k/ VOT continua averaged over 10 subjects. Each continuum consisted of eight members ranging in VOT from +15 msec to +50 msec. The continua were distinguished by the onset frequency of their first formants (208, 311 or 412 Hz) and by their frequency extent of the first formant transitions (0, 100, or 200 Hz).

		First Formant Onset Frequency (Hz)		
		208	311	412
First Formant	0	77.8	66.6	33.5
Transition	100	69.2	48.9	34.1
Extent (Hz)	200	60.4	52.9	41.5

A difference of 12.6 percent between any pair of means is sufficient for a posteriori significance at the $p < 0.01$ level.

TABLE 4 : Experiment II : Condition 2.

Mean phoneme boundaries in msec of VOT estimated by Probit Analysis for each of seven subjects (S) on each of three /g-k/ VOT continua. The continua were distinguished by the relative intensities of F_1 and F_2. F_1 varied through a 12 dB range across the three continua (-6 dB, 0 dB and +6 dB relative to F_2).

	Continuum: Relative Intensity of F_1:-		
	(1)	(2)	(3)
Subject	-6dB	0dB	+6dB
S1	36.39	34.56	33.76
S2	39.78	36.95	33.77
S3	39.65	39.56	39.54
S4	33.96	37.77	37.42
S5	33.92	37.26	35.98
S6	41.18	36.00	41.58
S7	27.88	32.22	36.05
MEANS	36.10	36.33	36.71

Four subjects claimed that more than 25 percent of the initial consonants were neither /g/ nor /k/. They heard some stimuli with long VOTs as initiated by palatal affricates (for example, /tʃi/). Their data were qualitatively similar, though more variable than those of the other subjects. The data of one experienced subject were noise free but will be mentioned no further as he only heard instances of /g/.

The numbers of /g/ responses afforded each of the 72 stimuli by each of the ten consistent subjects were examined in a three-way univariate analysis of variance with the factors:

a) subjects (10),
b) F_1 onset frequency (208, 311 or 412 Hz),
c) F_1 transition extent (0, 100 or 200 Hz) and
d) VOT (15, 20, 25, 30, 35, 40, 45 or 50 msec).

The effects of both the major independent variables and their interaction were significant (F_1 onset frequency: $F[2,18]=28.64$; $p < 0.01$. F_1 transition extent: $F[2,18]=11.38$; $p < 0.01$. Interaction: $F[4,36]=7.30$; $p < 0.01$). Post-hoc comparisons made according to the criteria recommended by Scheffe (1959) show that increasing F_1 onset frequency both from 208 Hz to 311 Hz, and from 311 Hz to 412 Hz, produced significant decreases in the percentage of /g/ responses ($p < 0.05$). Increasing F_1 transition extent from 0 Hz to either 100 Hz or 200 Hz, also produced a significant decrease in the percentage of voiced percepts ($p < 0.05$), but no systematic effect resulted from the increase from 100 Hz to 200 Hz of F_1 transition extent. The extent to which these results are manifest in individual comparisons may be examined in Table 3 where a difference of 12.6 percent between any pair of means is required for a posteriori significance at the $p < 0.01$ level.

Overall, the results show that increasing F_1 onset frequency reduces the proportion of voiced percepts independently of the characteristics of any following F_1 transition. The extent to which the presence of an F_1 transition also reduces the proportion of voiced percepts depends on its onset frequency. The effect is largest for onsets at 208 Hz, and diminishes to zero as the onset is raised to 412 Hz.

Control Condition 2

Two stimuli were added to the continuum used in Condition 1 in which F_1 had its onset at 311 Hz with 0 Hz F_1 transition extent. The extended continuum ranged from +10 msec to +55 msec of VOT. It was duplicated twice to create a total of three continua in which the level of F_1 relative to F_2 was +6 dB, 0 dB and -6 dB. Seven naive subjects listened to a randomization comprising ten instances of each of the 30 stimuli and indicated whether they perceived the initial consonant as /g/ or /k/. Table 4 shows their phoneme boundaries estimated by probit analysis (Finney, 1971).

These boundaries were examined in a two-way analysis of variance with the factors:

a) subjects (7) and
b) relative formant amplitude (+6 dB, 0 dB or -6 dB).

The effect of varying the relative amplitudes of F_1 and F_2 was not significant ($F[2,12]=0.093$). Although one subject did show a small increase in boundary position with increasing intensity of F_1, two others displayed the reverse pattern. Overall, variation of the relative intensities of F_1 and F_2 in these continua produced no systematic effect on the decision as to whether the initial stop was voiced or voiceless.

Discussion

The perceptual effects of varying F_1 onset frequency in Experiment I could have been mediated by those covariations in relative and overall formant amplitudes that the acoustic theory of speech production predicts, and that an OVE synthesizer produces. Had that been so, no effects should have resulted in Experiment II from varying the frequency of F_1 while holding its absolute and relative amplitude constant, but an appreciable effect should have resulted from varying its amplitude while holding its frequency constant. This was not the case. The opposite pattern was produced and confirms that the critical aspect of F_1 for the perceptual categorization of members of VOT continua is its perceived frequency at the onset of voicing, rather than its absolute or relative amplitude.

In Control Condition 1, the frequency extent of F_1 transition was varied while holding its onset frequency fixed. The results of this manipulation confirmed the second finding of Experiment I that a rising F_1 transition following the onset of voicing does not, in itself, increase the probability of a voiced percept. Transitions onsetting at 250Hz (in Experiment I) and at 208Hz and 311Hz (in Experiment II), significantly increased the probability of voiceless percepts. The physiological representations of the separation cue and the F_1 onset cue could both be influenced by whether voicing onset is accompanied by a rising, rather than a steady, F_1. If there were less energy in the critical band around the putative onset frequency of an F_1 transition than at the onset of a fixed frequency F_1, then the separation interval might be perceived as longer and the F_1 onset frequency as higher than their respective physical values. The data imply that the perceived onset of F_1 in these stimuli is determined by spectrotemporal integration over the duration of the first two or three pitch pulses, but that the dependency of F_1 onset registration on spectrotemporal integration decreases as physical onset frequencies increase from 200 Hz to 400 Hz.

Experiments I and II demonstrate that the perceived frequency of F_1 at the onset of voicing plays an identifiable role as a spectral parameter influencing the voiced-voiceless decision. They do not determine whether it is correct to impute to the frequency of the F_1 peak the entire burden of spectral influence or whether that influence derives from the distribution of energy in the spectrum including both F_1 and the higher formants. Lisker (1975) considered this possibility to be unlikely, although the perceived differences between his stimulus types can be economically summarized by expressing the spectral influence as the weighted sum of an effect of F_1 and an effect of F_2. A dependency of boundary location on the frequencies of

TABLES 5a, b, c: Obtained phoneme boundaries in msec of VOT and boundaries predicted by the equation:-

$$Vb = 58 - 100|(^2/_5)\mathrm{Log}(F_1^*/_{200}) + (^2/_3)\mathrm{Log}(F_2^*/_{1000})|$$

where:
Vb is the predicted boundary in msec of VOT,
F_1^* and F_2^* are the frequencies of the first and second formants at the onset of voicing.

TABLE 5a: The letters A, B, C, D and E identify five /g-k/ continua as in the original paper.

Continuum	F_1^*	F_2^*	Obtained	Predicted	Difference
A	540	1232	39	40	+1
B	769	1232	30	29	-1
C	386	1232	43	41	-2
D	286	1845	35	34	-1
E	412	2000	24	25	+1

Data from Lisker (1975).

TABLE 5b: Predictions are made for three of the seven continua.

Continuum	F_1^*	F_2^*	Obtained	Predicted	Difference
1	200	2098	34	37	+3
5	300	2158	27	29	+2
7	400	2194	23	23	0

Data from Experiment I: Condition 1.

TABLE 5c: Predictions are made for eight /da-ta/ continua differentiated by their F_1 transition durations (F_1T-Dur.).

Continuum F_1 T-Dur.	F_1^*	F_2^*	Obtained	Predicted	Difference
20	645	1200	21	32	+11
25	575	1235	22	34	+12
40	540	1320	30	33	+3
55	478	1375	34	34	0
70	452	1410	40	34	-6
85	427	1445	44	34	-10
100	400	1480	45	34	-11
115	375	1500	46	35	-11

Data from Lisker et al. (1975).

FIGURE 7

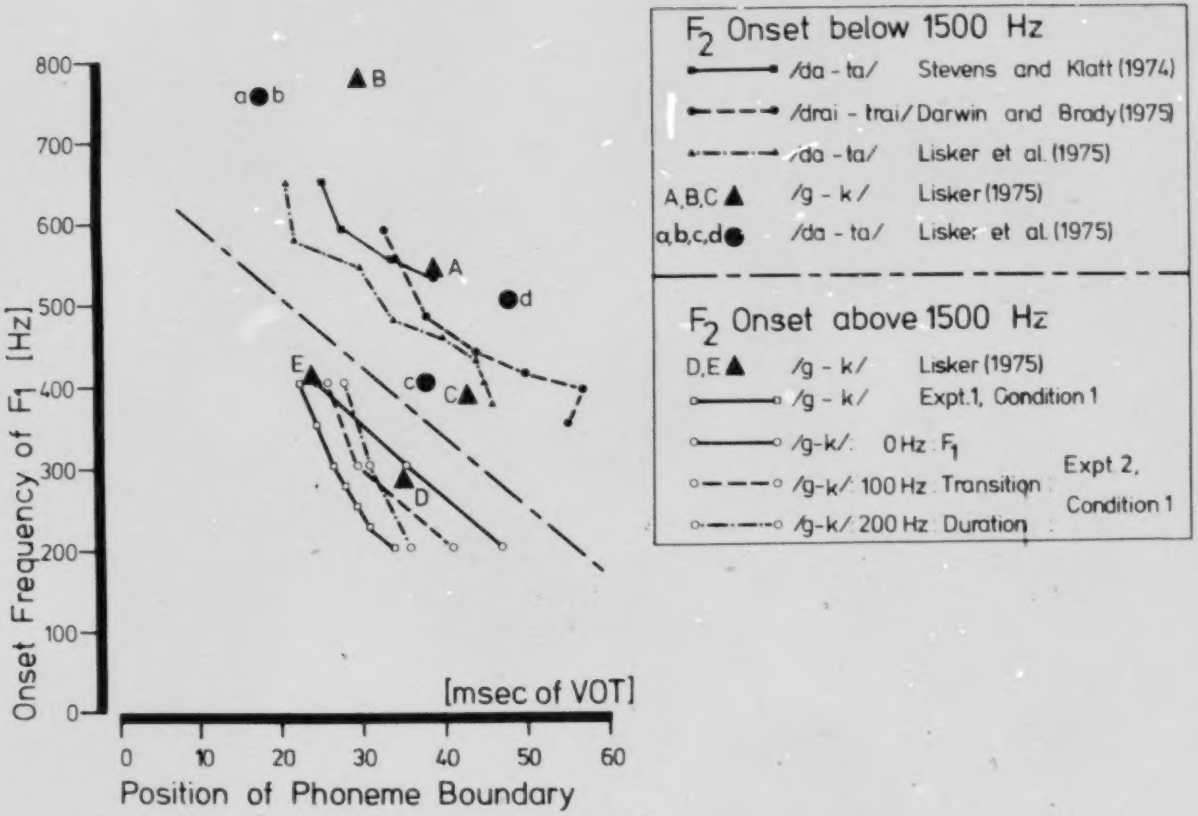


Figure 7: Plot of the position of the voicing boundary against the onset frequency of F_1 for the data sets indicated. The dotted line falling diagonally from left to right segregates the data according to the frequency of F_2 at the voicing boundary.

both F_1 and F_2 at the onset of voicing is economically expressed in the otherwise arbitrary formula:

$$Vb = 58 - 100\left[\frac{2}{5}\left|\text{Log}\left(\frac{F1^*}{200}\right)\right| + \frac{2}{3}\left|\text{Log}\left(\frac{F2^*}{1000}\right)\right|\right]$$

Where: Vb is the predicted voicing boundary of VOT in msec.

: F_1* and F_2* are the frequencies in Hz of the first and second formants at the onset of voicing.

The values of the constants were derived by trial and error to fit Lisker's (1975) data as shown in Table 5a. While the fit to Lisker's data is quite good, suggesting a role for F_2, and the expression adequately predicts the boundary positions observed here in Experiment I (shown in Table 5b), Table 5c shows that the equation fails to account for the data of Lisker, Liberman, Erickson and Dechovitz (1975).

Figure 7 displays a plot of obtained phoneme boundary location as a function of F_1 onset frequency for data reported in the present paper, and by Stevens and Klatt (1974), Lisker (1975), Lisker et al. (1975), and Darwin and Brady (1975). There are two important features of this display. First, the inverse relationship between the onset frequency of F_1 and the position of the voicing boundary demonstrated in the present experiments is equally apparent in the other sets of data plotted here. Second, despite the failure of the equation to describe the data of Lisker et al., the remaining data do justify the search for some description of spectral influences that includes the frequency of F_2 in addition to that of F_1. The dotted line in Figure 7 falling diagonally from left to right segregates the data according to the frequency of F_2 incorporated in the stimuli. Results obtained from stimuli in which F_2 was above 1500 Hz fall below this line, those in which F_2 was below 1500 Hz fall above the line. The pattern suggests that lowering the frequencies of both F_1 and F_2 can cause the voicing boundary to shift to longer VOTs. In addition, it appears that the more diffuse the spectrum the larger is the effect of varying F_1 onset frequency.

While this is one explanation for the pattern of data in Figure 7, it is also possible that the pattern reflects the effects of variations in voicing cues quite different from those considered here [see Klatt (1975) for a review], and the effects of different strategies for synthesis and the use of different groups of listeners. Resolution of these alternatives requires that the same group of listeners categorize the members of a set of VOT continua whose vocalic contexts are characterized by a range of F_2 frequencies in combination with a range of F_1 frequencies. This was done in Experiment III.

EXPERIMENT III

Stimuli and Procedure

Sixteen /d-t/ VOT continua were synthesized on the parallel formant synthesizer at Haskins Laboratories. The continua included identical synthesis control parameter specifications for F_3, F_0 and the overall and

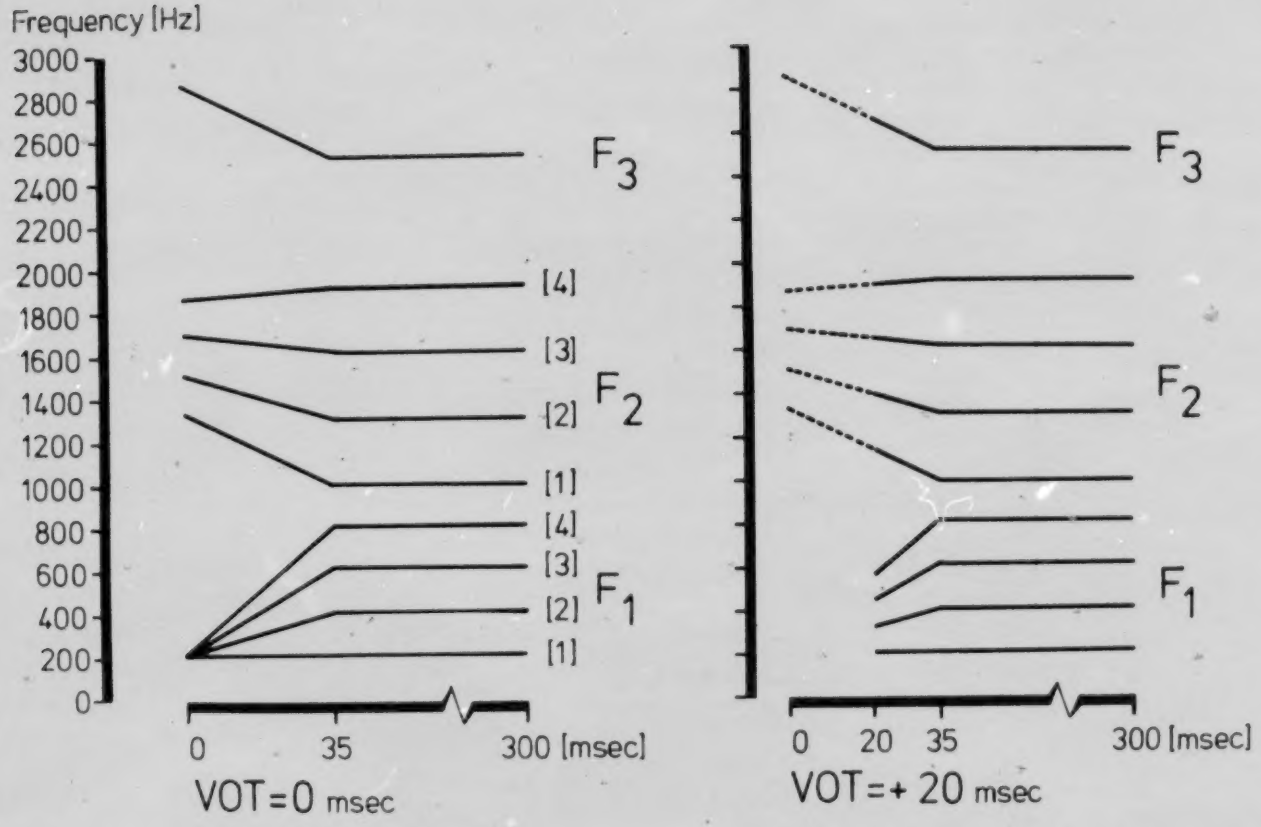


FIGURE 8

Figure 8: Schematic spectrograms showing the patterns of the first three formants for the stimuli used in Experiment III in exemplars with VOTs of 0 msec (left) and +20 msec (right). Sixteen VOT continua were created by combining each of four F_1 contours ([1]-[4]) with each of four F_2 contours ([1]-[4]). The stimuli included a 10 msec burst centered on 4000 Hz that is not shown.

individual formant amplitudes. They were distinguished only by differences in the frequency contours of their first and second formants. Sixteen continua were formed by combining each of four F_1 steady-state frequencies (208, 412, 614 and 819 Hz) with each of four F_2 steady-state frequencies (1001, 1306, 1611 and 1917 Hz). This range of formant frequencies includes vowels not found in the English vowel system. Transitions in F_1, F_2 and F_3 were linear in frequency/time over their duration of 35 msec. F_1 transitions rose from 208 Hz at stimulus onset to the appropriate steady-state. F_2 onset frequencies were computed so that the extrapolated trajectories of F_2 transitions originated at 1800 Hz 50 msec before syllable onset. The F_3 transition had its onset at 2861 Hz and fell to a steady-state at 2527 Hz. All stimuli included a fricated burst centered on 4000 Hz and lasting 10 msec from stimulus onset. Each stimulus was 300 msec in duration. Over the first 100 msec, the fundamental frequency was constant at 110 Hz. Figure 8 includes schematic displays of the formant parameter specifications of the stimuli.

Each continuum consisted of 10 members with VOTs of +5, +10 +15, +20, +25, +30, +35, +40, +45 and +50 msec formed by replacing periodic excitation with noise excitation in F_2 and F_3 and eliminating energy in F_1. The onset of pitch-pulsing was synchronized to the intended VOT in every stimulus.

Ten naive subjects listened to a randomization that included 10 instances of each of the 160 stimuli over Grason-Stadler TDH39-300Z headphones at a constant peak intensity of 85 dB SPL. They were instructed to make a forced-choice identification of the initial consonant of each stimulus as either /d/ or /t/ and to indicate their percept by writing 'D' or 'T'. In addition, subjects were instructed to mark with a '?' any response about which they were not confident.

Results

Despite being presented with a bizarre array of vowels, most subjects experienced little difficulty in performing the task. While four subjects did indicate that many of their responses to the members of the four continua with F_1 set to 200 Hz were guesses, no subject performed inconsistently with stimuli drawn from the other 12 continua.

The data were examined in three ways in different univariate analyses of variance. The first examined the sums of the numbers of 'D' responses made to each stimulus by each subject according to the four factors:

a) subjects (10),
b) F_1 steady-state (208, 412, 614 or 819 Hz),
c) F_2 steady-state (1001, 1306, 1611 or 1917 Hz) and
d) VOT (+5,+10,+15,+20,+25,+30,+35,+40,+45 or +50 msec).

Both the main effect of F_1 ($F[3,27]=26.27$; $p < 0.01$) and its interaction with VOT ($F[27,243]=13.27$; $p < 0.01$) were significant. Neither the main effect of F_2 ($F[3,27]=0.68$; $p > 0.2$), nor its interaction with VOT were significant. The data provided by the six subjects who performed consistently on all sixteen continua were examined in probit analyses that fitted

ogives to the data from each subject for each continuum. Two parameters were estimated for each ogive: the physical stimulus value corresponding to the p 0.5 point on the psychometric function and the slope of the probit regression. The first parameter is an estimate of the phoneme boundary. The second varies directly with the standard deviation of the psychometric function underlying the test continuum and hence reflects the slope of the identification function at the boundary. The two parameters were examined in separate analyses with the factors:

a) subjects (6),
b) F_1 steady-state (208, 412, 616 or 818 Hz) and
c) F_2 steady-state (1001, 1306, 1611 or 1917 Hz).

Analysis of the 50 percent intercepts that correspond to the phoneme boundary, showed a significant effect of F_1 ($F[3,15]=35.95$; $p < 0.001$), nonsignificant effects of F_2 ($F[3,15]=0.84$; $p > 0.2$), and no F_1xF_2 interaction ($F[9,45]=0.48$; $p > 0.2$). Analysis of the boundary slopes also showed a significant effect of F_1 ($F[3,15]=5.00$; $p < 0.025$), nonsignificant effects of F_2 ($F[3,15]=0.05$; $p > 0.2$), and no F_1xF_2 interaction ($F[9,45]=1.93$; $p > 0.1$).

These results may be assessed in relation to the plots in Figure 9 where boundary position is plotted against the steady-state frequency of F_2 for each value of F_1 steady-state frequency. Only data provided by the six subjects who performed consistently on all sixteen continua are represented. The plots corresponding to each value of F_1 onset frequency are horizontal, illustrating the lack of any dependency of boundary position on F_2 onset frequency. Means obtained by averaging over these subjects are tabulated in Table 6 which shows that as the F_1 steady-state increases in frequency, two things do happen: phoneme boundaries shift to shorter VOTs and the slopes of the probit regressions, and hence of the identification functions at the boundary, become steeper.

Discussion

It is clear that, overall, the perceived frequency of F_2 at the onset of voicing plays an insignificant role in determining how listeners categorize the members of /d-t/ VOT continua as voiced or voiceless[3]. It is unlikely that the absence of an F_2 effect here, as contrasted with Lisker's (1975) data, results from our use of the alveolar rather than the velar place of production. Comparison of the data from Experiment III with that plotted in Figure 7 shows our velar and alveolar data to correspond quite precisely. While Lisker's (1975) data remain anomalous, the present result is congruent with two earlier observations. Summerfield (1974a) varied the durations of syllable-initial F_1, F_2 and F_3 transitions in the members of /ga-ka/ and /gi-ki/ VOT continua. This produced a systematic change in the position of the

[3]However, see Draper and Haggard (1974), Sawusch and Pisoni (1974), and Repp (1976) for discussions of effects on the perception of place and voicing deriving from the microstructure of F_2 and F_3 transitions, as opposed to the macroscopic effect sought here.

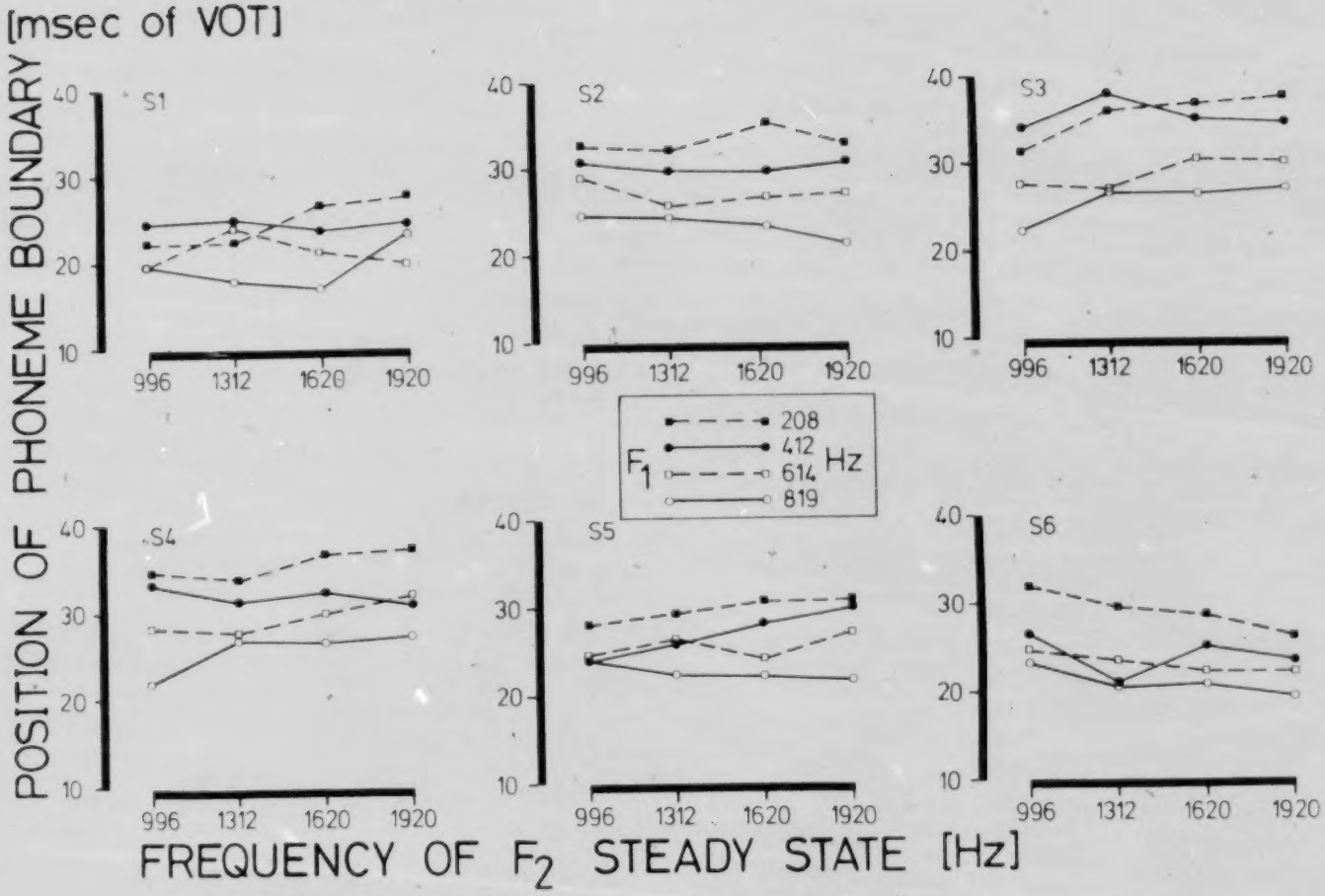


FIGURE 9

Figure 9: Results of Experiment III for six individual subjects who performed consistently on all sixteen continua. For each subject four empirical functions relate the position of the voicing boundary (estimated by probits) to the frequency of the F_2 steady state for each of four values of F_1 steady state. The functions are essentially horizontal showing no dependency of the position of the voicing boundary on the spectral characteristics of F_2.

TABLE 6: Experiment III.

Phoneme boundary positions in msec of VOT averaged over six subjects whose data were internally consistent on all sixteen continua.
The continua were distinguished by the frequency of their F_2 steady-states (1001 Hz, 1306 Hz, 1611 Hz or 1917 Hz) and the frequency of their F_1 steady-states (208 Hz, 412 Hz, 614 Hz or 819 Hz).
Four values are indicated for each continuum. The first is the position of the average phoneme boundary in msec of VOT [PB]. The second is the average slope of the Probit regression line [SL]. Its units are (Probit of [+voiced] responses)/(ms.).
The third and fourth values are the frequencies of the first and second formants at the mean phoneme boundary locations (F_1* and F_2*).

	Continua with F_1=208 and F_1=412 Hz:- F_1 steady-state (Hz), F_2 steady-state (Hz) :-							
[F_1]	208	208	208	208	412	412	412	412
[F_2]	1001	1306	1611	1917	1001	1306	1611	1917
MEANS								
[PB]	30.09	30.47	32.13	31.94	28.83	28.34	28.86	28.92
[SL]	-0.142	-0.158	-0.167	-0.138	-0.187	-0.182	-0.174	-0.167
[F_1*]	208	208	208	208	381	381	381	381
[F_2*]	1001	1306	1611	1917	1077	1382	1611	1917

	Continua with F_1=616 Hz and F_1=818 Hz:- F_1 steady-state (Hz), F_2 steady-state (Hz):-							
[F_1]	616	616	616	616	818	818	818	818
[F_2]	1001	1306	1611	1917	1001	1306	1611	1917
MEANS								
[PB]	25.59	25.36	25.73	26.34	22.84	22.24	22.33	22.99
[SL]	-0.188	-0.172	-0.158	-0.169	-0.195	-0.185	-0.195	-0.227
[F_1*]	535	535	535	535	611	611	611	611
[F_2*]	1077	1382	1611	1917	1077	1382	1611	1917

phoneme boundary in /a/-context, where there was an extreme F_1 transition whose onset frequency at any given VOT varied with transition duration. However, there was no effect in /i/-context, where, despite a negligible F_1 transition, there were appreciable transitions in F_2 and F_3 whose onset frequencies did vary. Lisker et al. (1975) varied the durations of the F_2 and F_3 transitions independently of that in F_1 in the members of a /da-ta/ continuum. Systematic changes in the position of the voicing boundary resulted from manipulations of F_1, but not from those of F_2 and F_3. The results of Experiment III augment these earlier findings. They demonstrate that the major spectral influence on the perception of stop-voicing resides in F_1 and is not distributed throughout the entire spectrum. Perceptual behavior is explained in terms of the direct acoustic effects of particular vocalic environments on the voicing cues without the invocation of feedback from the phonetic identification of the vowel.

For each steady-state frequency of F_2 used in Experiment III, the empirical function relating the position of the phoneme boundary to the onset frequency of F_1, if plotted in Figure 7, would cross the dotted line that has been purported to segregate results according to the frequency of F_2 incorporated in the stimuli. Clearly, a different rationale for the pattern of data in Figure 7 is required. The explanation may be found in the observation that the different data sets displayed derived from stimuli with different overall durations. The stimuli of Lisker et al. (1975) and Darwin and Brady (1975) had durations of 600 msec, while those of Lisker (1975) were 450 msec, and those used in the present experiments were 300 msec in duration. Summerfield and Haggard (1972) observed that increasing the duration of the steady-state portion of a CV syllable with a fixed VOT increased the probability that the initial consonant would be perceived as voiced. They argued that this finding demonstrated perceptual sensitivity to acoustic covariants of speech rate. We have replicated this finding and found that an increase from 90 msec to 310 msec in the duration of the vowel in the members of a /biz-piz/ continuum shifts the position of the voicing boundary by about 7 msec. A simple mechanism that could simulate this effect would scale the duration of the separation interval in a stimulus in relation to the total duration of the syllable, combine the scaled duration with measures of other pertinent cues, and compare the combined cue-value with a criterion value to determine the value of the voicing feature. If the effect of manipulating the physical value of another cue, for example, F_1 onset frequency, were assessed by measuring changes in the position of the voicing boundary expressed in terms of the physical value of the separation interval, then the measured effect would increase as the total duration of the stimulus increased. The relation between the present data and that of Lisker et al. (1975) and Darwin and Brady (1975) is congruent with this rationale; larger effects of F_1 onset frequency variation were produced by these authors' 600 msec stimuli than by our 300 msec stimuli. This explanation remains to be tested and does not account for the patterns of Stevens and Klatt's (1974) and Lisker's (1975) data; those data remain anomalous.

GENERAL DISCUSSION

Trading Relationships in Production and Perception

These results identify the perceived frequency of the first formant at the onset of voicing as the critical spectral parameter influencing the perceptual categorization of members of VOT continua. They have shown that a larger value of the separation interval, the purely temporal component of VOT, is required for the perception of a voiceless stop when F_1 has a low onset frequency (indicating greater vocal tract constriction) and vice versa. This trading realtionship corresponds elegantly with one in production.

In production, oral release gestures of differing extents made by the same articulators nevertheless tend to require the same length of time (for example, Kent and Moll, 1969; Perkell, 1969). It is observed that VOT varies inversely with both the rate at which the oral release gesture is made and with the degree of vocal tract constriction required by the phoneme following the stop. Thus, longer VOTs characterize velar stops, compared to alveolars, compared to bilabials (Lisker and Abramson, 1964); VOTs tend to be longer before the vowel /i/ than before /a/ (Klatt, 1975; Summerfield, 1975a); VOTs are longer in stop+/r/+vowel and stop+/l/+vowel environments than in stop+vowel environments (Klatt, 1975).[4] It is not entirely clear why this relationship occurs in production. A relatively constricted vocal tract both increases the acoustic load on the glottal source (Flanagan and Landgraf, 1968), and may also retard the attainment of the transglottal pressure drop necessary for vocal cord vibration (van den Berg, 1968). Klatt (1975) points out, however, that passive aerodynamics can only contribute to variations in VOT observed in productions of voiced stops, since in voiceless productions the supraglottal pressure established during the occlusive phase is entirely dissipated during the fricative portion of the stop-release and is at

---

[4] L. Lisker (1961) Voicing lag in clusters of stop plus /r/. Haskins Laboratories Final Report on Speech Research and Instrumentation (unpublished). Lisker reports VOTs measured in syllable-initial voiceless stops preceding a vowel and preceding /r/+vowel as follows:

/p/: +61 msec, /pr/: +89 msec;
/t/: +64 msec, /tr/: +110 msec;
/k/: +77 msec, /kr/: +107 msec.

Klatt (1975) reports similar data for voiceless plosives and the following data for voiced plosives:

/b/: +7 msec, /br/: +12 msec;
/d/: +14 msec, /dr/: +29 msec;
/g/: +23 msec, /gr/: +32 msec.

It is noteworthy that a putatively voiced, syllable-initial /gr/ can be characterized by a VOT almost twice as large as the simultaneity threshold (Hirsh, 1959) that has been invoked as a psychoacoustic basis for the voicing distinction in English (for example, Miller et al., 1976; Pisoni, in press).

atmospheric level at the time when vocal cord adduction is initiated. He suggests that, to offset the inherently low frequency of F_1 when stops are produced before a close vowel or a lateral, the timing of glottal adduction relative to oral release could be actively delayed.[5] It is fairly parsimonious to postulate such learned compensation in production. Perceptual sensitivity to the summed cue values of separation interval and F_1 onset frequency is already required, whatever the habits of production may be. By pooling measures of these two cues at a low level, the noninvariance problem for perception is reduced. This perceptual summation should apply equally in the speaker's perception of his own productions. As a quid pro quo, production could be expected to develop vowel contingent modifications to delay adduction in order to permit a general criterion value of the summed measure to characterize phoneme boundaries in most circumstances. Possibly, small passive aerodynamic effects of the adjacent vowel upon voicing onset occur in unstressed syllables, while larger delays result from controlled adduction delay in stressed syllables.

The identification of the role of F_1 onset frequency permits the rationalization of a group of previously reported results. In Figure 10, four F_1 transition contours that might be incorporated in the members of synthetic VOT continua are schematized. Transitions [a] and [b] differ in duration, while transitions [b] and [c] differ in spectral extent. Contour [d] evinces no transition. Were voicing to onset at time T_1 msec, F_1 onset frequencies of F_a, F_b, F_c and F_d Hz would result. The diagram exemplifies, as Lisker et al. (1975) have emphasized, that variation in either the temporal duration or the frequency extent of an F_1 transition results in covariation of F_1 onset frequency at any given VOT. Thus, effects previously attributed to F_1 transitions following experimental manipulation of either transition duration (Stevens and Klatt, 1974; Summerfield, 1974a), or frequency extent (Summerfield and Haggard, 1974), where the F_1 steady state was fixed, are more appropriately ascribed to variation in F_1 onset frequency. Similarly, phoneme boundaries on VOT continua involving the vowel /i/ (with a low frequency F_1 in the vowel and hence little or no F_1 transition) fall at longer VOTs than do those on continua with the vowel /a/ (with a high frequency F_1 in the vowel and a potentially extensive F_1 transition) (Cooper, 1974; Summerfield, 1974a; 1975b); that finding is rationally explained by the necessarily lower F_1 onset frequency in /i/-context. (Compare contours [b] and [d] in Figure 10.) These results would be paradoxical if the transition were considered to be a cue to voicedness; the paradox led Summerfield and

---

[5] We and our colleague Peter Bailey have recently measured periods of devoicing and VOTs in productions of /p/, /t/ and /k/ before /i/, /a/, /ri/ and /ra/ in bisyllables such as /bəpri/. Total periods of devoicing (that is, the time from the disappearence of periodicity in the waveform at approximately the moment of stop closure to its reemergence at voicing onset), tend to be more invariant than either the period of devoicing preceding oral release or the VOT itself. Possibly observed covariations of VOT with the degree of vocal tract constriction required by the following phoneme reflect an active process in which it is the moment of oral release that is varied within a fixed time-frame of adduction-abduction.

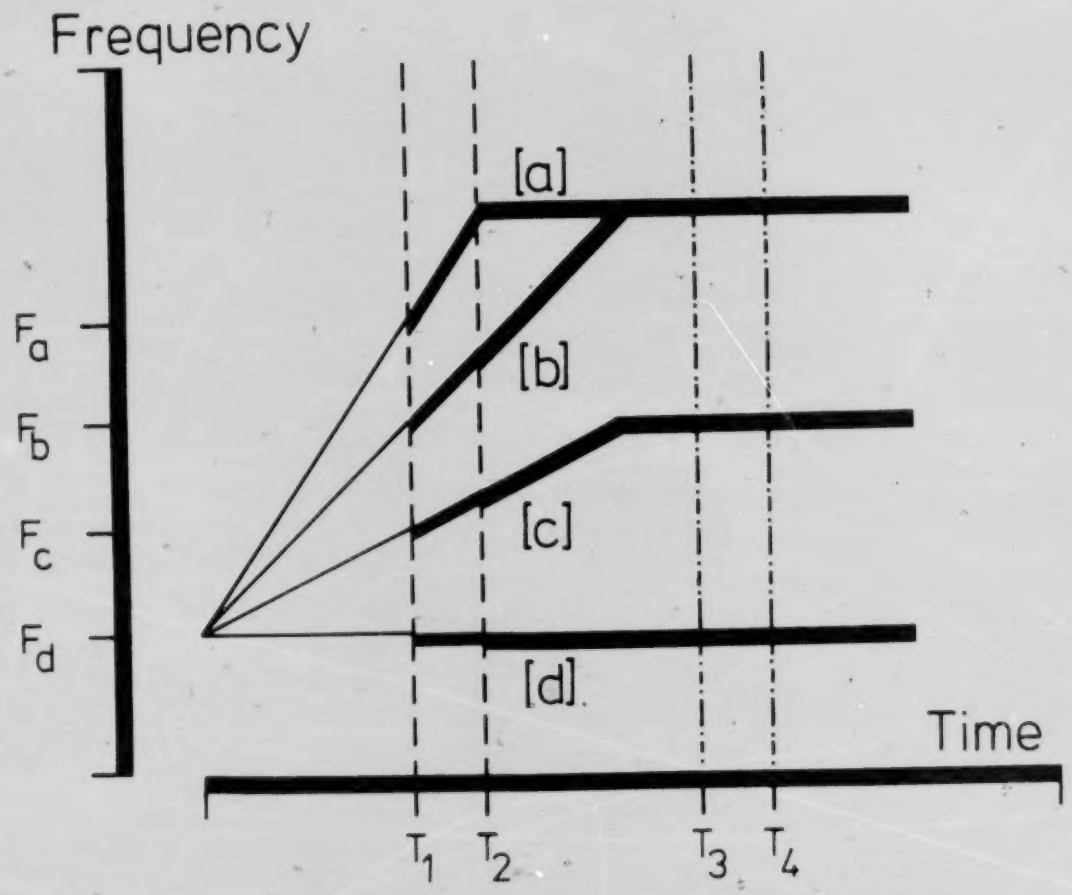


Figure 10: Schematic descriptions of four syllable-initial first formant contours ([a], [b], [c], [d]) which could be incorporated in the members of different VOT continua. Were voicing to onset at time T_1, first formant onset frequencies of F_a, F_b, F_c and F_d Hz would result.

Haggard (1974) to consider a possibility that they otherwise acknowledged to be unparsimonious, namely that the perceptual weightings of measures of the temporal and spectral aspects of VOT might be conditioned by vocalic context. With F_1 onset frequency identified as the critical spectral parameter, there is no need for such feedback, and the voicing decision may be reached without reference to the category of phoneme following the stop. (See also Darwin and Brady, 1975.) Further methodological implications are reviewed in a following section.

The results obtained here may reflect the effects of another, less influential spectral parameter, in addition to F_1 onset frequency. The schematic displays in Figures 1, 2, 5 and 8 show that the constraints that were applied to the acoustic structure of the stimuli necessarily resulted in covariation of the frequency of the F_1 steady-state with, in different conditions, either the onset frequency of F_1, or the extent of the F_1 transition. Increases in both these latter variables raised the probability that a stop-consonant characterized by a particular VOT would be perceived as voiceless. Thus, the results exhibit a correlation between the frequency of the F_1 steady-state and the probability of a voiceless percept. Experiment I showed that there is not a strong causative relationship between the two. However, the results do not eliminate the possibility that there may be some influence. Stevens and House (1963) noted that the contour of F_1 in the vocalic portions of natural CV syllables is lower in frequency following voiced, as opposed to voiceless, stops--reflecting the increase in vocal tract length that results from the lower position of the larynx in voiced productions, (for example, Ewan and Krones, 1974). This aspect of articulatory behavior increases the spectral difference in F_1 at voicing onset between voiced and voiceless productions. It remains to be determined whether an additional perceptual effect derives from the coarticulated variation in the F_1 steady-state.

First Formant Transitions and First Formant Onsets

The failure of an F_1 transition to cue voicing in adults raises doubts about Stevens and Klatt's (1974) suggestion as to its perceptual primacy for the perception of voicing contrasts in infants. Such wariness is reinforced by two recent findings. First, demonstrations of the categorical perception of the members of continua formed by varying the relative onset times of noise and buzz segments (Miller, Pastore, Wier, Kelley and Dooling, 1976) and pairs of sine waves (Pisoni, in press) have confirmed Hirsh's claim (Hirsh, 1959; Hirsh and Sherrick, 1961) that a natural psychoacoustic boundary between the perception of successive and simultaneous coterminous acoustic events occurs at a temporal offset of about 17 msec. Although as the results of the present experiments show, the perception of voicing contrasts involves the registration of the spectral concommitants of the interval between release and voicing onset, psychoacoustic considerations may well dictate why a temporal interval is the basis of the voicing distinction in general (whether positive or negative values of VOT are involved), and why in particular many of the world's languages place a category boundary between VOTs of 0 and +40 msec. The second difficulty for the supposed primacy of transitions comes from a developmental study by Simon (1974). He showed that children older than eight years do not categorize any members of a 'Goat-Coat' VOT continuum as initiated by [g], unless they contain a low F_1 onset

frequency. Children younger than five years, on the other hand, indicate that they have perceived [g] in the absence of the spectral cue and appear to be primarily sensitive to variation in the temporal cue. These results support Lisker's assertion of the primacy of the temporal aspect of VOT and suggest that it is the ability to detect the spectral cue that is learnt.

At present, it is not clear whether infants' behavior in discriminating members of VOT continua (c.f. Eimas et al., 1971; Streeter, 1976) represents a psychoacoustic ability to distinguish successive from simultaneous acoustic events, or a phonetic ability to distinguish voiced from voiceless stops (Pisoni, in press). The alternatives could be dissociated by experimenting with VOT continua (for example, /gri-kri/) on which the phoneme boundary, by virtue of a low F_1 onset frequency, occurred at a considerably longer VOT than the simultaneity-succesivity threshold. Would infants discriminate better across the psychoacoustic boundary, the phonetic boundary, or both?[6]

Implications for Studies Using Stimuli Drawn from VOT Continua

The demonstration that the temporal and spectral components of VOT may be traded for one another and that, by implication, each possesses perceptual potency in cueing the voicing distinction, has methodological import for studies whose stimuli are drawn from VOT continua.

Where F_1 transition duration is held unnaturally constant across continua that represent articulations in which it would normally vary, the positions of phoneme boundaries should not vary. Darwin and Brady (1975) synthesized /de-te/ and /dri-tri/ continua with identical parametric specifications of F_1. The perceptual identification functions for the two continua differed slightly, but in the reverse direction from that to be expected if the boundary locations were determined by phonetic class: boundaries on the /dri-tri/ continuum occurred at shorter VOTs than those on the /de-te/ continuum. Lisker et al. (1975) synthesized /ba-pa/, /da-ta/ and /ga-ka/ continua with identical transition specifications for F_1. Boundaries on these three continua coincided, in contrast to those obtained in Lisker and Abramson's original (1967) study where the duration of the F_1 transition covaried naturally with place of production.[7]

If VOT continua involve cutback of the duration/frequency-extent of an F_1 transition, then variation in VOT over the duration of this transition (for example, between times T_1 and T_2 in Figure 10) will alter the physical

---

[6]The value of this test would be nullified if the psychoacoustic simultaneity threshold varied as a function of the frequency of the lower spectral component of the stimulus. This possibility is currently under investigation.

[7]A small place-voicing correlation, equivalent to a shift in the VOT boundary of about plus or minus 2 msec, remains even when all acoustic differences between stimuli are neutralized (see Draper and Haggard, 1974; Sawusch and Pisoni, 1974; Repp, 1976; Miller, in press).

values of both cues. Equivalent variation beyond the end of the transition (for example, between times T_3 and T_4), or on continua not involving an F_1 transition (for example, between either T_1 and T_2 or between T_3 and T_4 on contour [d]), will only vary the value of the separation cue. If, as the results of the present experiments suggest, the decision as to the value of the voicing feature may be represented as being based on a combination of analogue measures of these two cues and others (Hoffman, 1958; Haggard, 1974; Summerfield, 1974b)[8], then the perceptual effect of a particular change in VOT will depend upon the magnitude of the change in the combined value of the cues that it produces. A VOT shift that changes the physical values, and hence the perceptual measures, of both cues should produce a larger perceptual effect than should one that only varies the value of the separation cue. It is likely, in addition, that the perceptual scaling of the temporal separation component of VOT for values greater than the simultaneity-successivity threshold approximates Weber's Law (Abel, 1972; Miller et al., 1976). As a result of both these factors, the perceptual effect of a change in VOT of fixed size should diminish as the absolute VOT on which that change is centered increases. The perceptual consequences of the two factors have not been dissociated, although effects have been observed that reflect their joint operation. Pisoni and Lazarus (1974) carried out 4IAX discrimination tests of members of a /ba-pa/ continuum involving syllable-initial formant transitions of 50 msec duration. They noted that discrimination of stimuli differing in VOT by 20 msec was more accurate in the voiced range of VOTs from 0 to 40 msec, where the physical values of both cues were changing, than in the voiceless range above 40 msec. Similarly, Summerfield (1975c) measured phoneme boundary widths, defined as the difference between the VOTs corresponding to 25 percent and 75 percent voiced responses for each of eight subjects on a /ga-ka/ continuum that was synthesized with an extensive rising F_1 transition of 60 msec duration and on a /gi-ki/ continuum that was synthesized with no F_1 transition. Boundary width, in this definition, relates inversely to discrimination in the boundary region and should reflect the rate of change of the combined value of the two cues at the boundary. Mean phoneme boundaries occurred at +29.0 msec in /a/-context and at +41.6 msec in /i/-context. Mean boundary widths were 6.6 msec in /a/-context and 10.5 msec in /i/-context. Each of the eight subjects displayed larger boundary widths on the /gi-ki/ continuum than on the /ga-ka/ continuum. Similarly, estimates of the slope of the psychometric functions underlying the continua in Experiment III decreased significantly as mean phoneme boundary location increased. In all these studies, discrimination of VOT differences was best (a) at shorter as opposed to longer VOTs, and (b) when the change in VOT to be discriminated varied both the separation interval and the onset frequency of the first formant.

An implication of these observations is that the size of the change in the position of the phoneme boundary on a VOT continuum induced by a given difference in some contextual variable will be greatest when the induced

---

[8] A. Q. Summerfield (1975c) Information-processing analyses of perceptual adjustments to source and context variables in speech. Doctoral Dissertation, The Queen's University of Belfast (unpublished).

change occurs at a large mean VOT and only varies the duration of the separation cue. It will be smallest when the change occurs at short VOTs and varies both the onset frequency of F_1 and the duration of the separation cue. Summerfield (1975b) measured the size of shifts in the phoneme boundary on VOT continua caused by variation in the syllabic rate of phrases that introduced test syllables drawn from the continua. On continua synthesized with the vowel /i/ (where F_1 was low in frequency and there was only a small F_1 transition), phoneme boundaries fell at longer VOTs and larger phoneme boundary shifts were measured, than on continua with the vowel /a/ (where there was an extensive F_1 transition). The observations confirm the above deductions concerning discriminability and lend force to recent warnings by Abramson (1976) that the VOT dimension, though a simple temporal continuum when viewed in articulatory terms, involves variation in a complex set of acoustic parameters whose relative availability is a function of both absolute VOT and phonetic context. The interpretation of data obtained with stimuli drawn from such continua is only valid if it takes this complexity into account.

SUMMARY AND CONCLUSIONS

The experiments reported here permit two conclusions: (1) The perceived onset frequency of F_1 is the critical spectral parameter included in the repertoire of cues to the voicing decision for syllable-initial prestressed stop-consonants in English. The spectral influence derives only from F_1, not from the spectrum comprising F_1 and the higher formants. (2) A periodically excited, rising first formant transition is not, per se, a positive cue to voicing when its onset frequency is controlled.[9]

In perception, the temporal separation component of VOT and the F_1 onset frequency component may be traded one for the other: the lower the frequency of F_1 at the onset of voicing, the longer the separation interval required to produce a voiceless percept. This trading relationship parallels one in production where VOT varies inversely with the degree of vocal tract constriction, and hence with the frequency of F_1, required by the phoneme following the stop-consonant.

The greater role of F_1 onset frequency than or F_1 transition here does not imply that transition characteristics are never important in speech perception. A rising first formant at the onset of a pattern of formant frequencies signals an obstruent articulation and is more likely to predispose a consonantal percept, than is a fixed-frequency transitionless first

---

[9]Not all aspects of the present results are entirely novel. Liberman, Delattre and Cooper (1958) noted that cutting back F_1 changed the values of two correlated variables: the onset time of F_1 relative to F_2 and F_3; and the onset frequency of F_1. They demonstrated that relative onset time has perceptual significance independent of onset frequency. Whether F_1 onset frequency had independent perceptual significance was not reported at the time. The intervening years have enabled us to bring more sophisticated synthesis, psychophysical methods, and both psychological and articulatory interpretations to the classical problem of specifying the cues.

formant onsetting at the same frequency. Such a rapid spectral change need not be confined to the spectrum above 1 kHz as Stevens (1975) suggests. It is something to which an F_1 transition, relieved of the burden of characterizing (+voiced), contributes.

REFERENCES

Abel, S. M. (1972) Discrimination of temporal gaps. J. Acoust. Soc. Am. 52, 519-524.

Abramson, A. S. (1976) Laryngeal timing and consonant distinctions. Haskins Laboratories Status Report on Speech Research SR-47, 105-112.

Berg, J. W. van den. (1968) Mechanism of the larynx and the laryngeal vibrations, in Manual of Phonetics, ed. by B. Malmberg. (London: North-Holland), pp. 278-308.

Cooper, W. E. (1974) Contingent feature analysis in speech perception. Percept. Psychophy. 16, 201-204.

Darwin, C. J. and S. A. Brady. (1975) Voicing and juncture in stop-/r/ clusters. J. Acoust. Soc. Am. 57, S24(A).

Draper, G. (1973) SPEX: a system to run speech perception experiments. Proceedings of the 9th DECUS Europe Seminar. (Maynard, Mass.: Digital Equipment Users Society), pp. 89-93.

Draper, G. and M. P. Haggard. (1974) Facts and artifacts in feature interdependence. In Preprints of the Stockholm Speech Communication Seminar, vol. 3, ed. by G. Fant (Uppsala: Almqvist and Wiksell), pp. 67-75.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Ewan, W. G. and R. Krones. (1974) Measuring larynx movement using the thyroumbrometer. J. Phonetics 2, 327-336.

Fant, G. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Ferguson, G. A. (1966) Statistical Analysis in Psychology and Education. (London: McGraw-Hill).

Finney, D. J. (1971) Probit Analysis. (New York: Cambridge University Press).

Flanagan, J. L. and L. Landgraf. (1968) Self-oscillating source for vocal-tract synthesisers. IEEE Trans. Audio. Electroacoust. AU-16, 57-84.

Haggard. M. P. (1974) The perception of speech. In The Physics and Psychology of Hearing, edited by S. Gerber (Philadelphia: W. B. Saunders).

Hirsh, I. J. (1959) Auditory perception of temporal order. J. Acoust. Soc. Am. 31, 759-767.

Hirsh, I. J. and C. E. Sherrick. (1961) Perceived order in different sense modalities. J. Exper. Psychol. 62, 423-432.

Hoffman, H. S. (1958) Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am. 30, 1035-1041.

Kent, R. D. and K. L. Moll. (1969) Vocal tract characteristics of the stop cognates. J. Acoust. Soc. Am. 46, 1549-1555.

Klatt, D. H. (1975) Voice onset time, frication and aspiration in word-initial consonant clusters. J. Speech Hearing Res. 18, 686-706.

Liberman, A. M., P. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position.

Lang. Speech 1, 153-167.
Lisker, L. (1975) Is it VOT or a first-formant transition detector? J. Acoust. Soc. Am. 57, 1547-1551 (L).
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 324-422.
Lisker, L. and A. S. Abramson. (1967) The voicing dimension: some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague 1967. (Academia: Prague, 1970), pp. 563-567.
Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.
Lisker. L., A. M. Liberman, D. M. Erickson, and D. Dechovitz. (1975) On pushing the voice onset time boundary about. Haskins Laboratories Status Report on Speech Research SR-42/43, 257-264.
Mattingly, I. G. (1968) Synthesis by Rule of General American English. Doctoral Dissertation, Yale University. (Supplement to Haskins Laboratories Status Report on Speech Research.)
Miller, J. L. (in press) The perception of voicing and place of articulation in initial stop consonants: Evidence for the nonindependence of feature processing. J. Sp. Hear. Res.
Miller, J. D., R. E. Pastore, C. C. Wier, W. J. Kelly, and R. J. Dooling. (1976) Discrimination and labelling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Am. 60, 410-417.
Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge: MIT Press).
Pisoni, D. B. (in press) Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. J. Acoust. Soc. Am.
Pisoni, D. B. and J. M. Lazarus. (1974) Categorical and non-categorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Am. 55, 328-333.
Repp, B. H. (1976) The voicing boundary as a function of F_2 and F_3 transitions and fundamental frequency. J. Acoust. Soc. Am. 60, S91(A).
Sawusch, J. R. and D. B. Pisoni. (1974) On the identification of place and voicing features in stop consonants. J. Phonetics 2, 181-194.
Scheffe, H. (1959) The Analysis of Variance. (New York: Wiley).
Simon, C. (1974) Some aspects of the development of speech production and perception in young children. In Preprints of the 1974 Stockholm Speech Communication Seminar, vol. 4, ed. by G. Fant. (Uppsala: Almqvist and Wiksell), 7-14.
Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and the Perception of Speech, ed. by G. Fant and M. A. A. Tatham (London: Academic), pp. 303-330.
Stevens, K. N. and A. S. House. (1963) Peturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.
Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.
Streeter, L. A. (1976) Language perception of 2-month old infants shows effects of both innate mechanisms and experience. Nature 259, 39-41.

Summerfield, A. Q. (1974a) Processing of cues and contexts in the perception of voicing contrasts. In Preprints of the 1974 Stockholm Speech Communication Seminar, vol. 3, ed. by G. Fant. (Uppsala: Almqvist and Wiksell), pp. 77-86.

Summerfield, A. Q. (1974b) Towards a detailed model for the perception of voicing contrasts. Speech Perception. (Progress Report, Department of Psychology, The Queen's University of Belfast.) no. 3, 1-26.

Summerfield, A. Q. (1975a) How a detailed account of segmental perception depends on prosody and vice-versa. In Structure and Process in Speech Perception, ed. by A. Cohen and S. G. Nooteboom. (New York: Springer-Verlag), pp. 51-68.

Summerfield, A. Q. (1975b) Cues, contexts and complications in the perception of voicing contrasts. Speech Perception. (Progress Report, Department of Psychology, The Queen's University of Belfast.), no. 4., 99-130.

Summerfield, A. Q. and M. P. Haggard. (1972) Speech rate effects in the perception of stop voicing. Speech Synthesis and Perception. (Progress Report, Psychological Laboratory, University of Cambridge.), no. 6, 1-12.

Summerfield, A. Q. and M. P. Haggard. (1974) Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. J. Phonetics 2, 279-294.

Taylor, M. M. and C. D. Creelman. (1967) PEST: Efficient estimates on probability functions. J. Acoust. Soc. Am. 41, 782-787.

Perceptual Integration and Selective Attention in Speech Perception: Further Experiments on Intervocalic Stop Consonants

Bruno H. Repp

ABSTRACT

Three experiments on the perceptual interaction between implosive and explosive formant transitions of intervocalic stop consonants were conducted using synthetic VCV utterances. Experiment I demonstrates that implosive transitions are difficult to perceive correctly when followed by a steady-state vowel after a short silent interval (closure). Thus, perception of the stop is interfered with even when no conflicting explosive transitions follow the closure period. The same experiment also shows that VCV stimuli in which the implosive transitions are followed by conflicting explosive transitions are difficult to discriminate from stimuli in which the implosive transitions are phonetically compatible with the explosive transitions or absent altogether, as long as the closure duration is sufficiently short. Thus, the interference effect is as pronounced in terms of discrimination performance as it is in identification. Experiment II, a reaction-time (RT) task, replicates the finding that "same" judgments about the medial consonants in two successive VCV utterances are faster and more accurate when the final vowels are the same than when they are different. Eliminating the explosive transitions does not reduce the effect, not even at relatively long closure durations, which indicates a general perceptual integration effect that is not mediated by the acoustic covariation of explosive transitions with the final vowel. The data suggest that, in addition to complete stimulus identity -- which apparently is detected at a prephonetic, holistic stage of processing -- equality of overall stimulus structure (VC vs. VCV) facilitates "same" judgments. The size of the perceptual units compared seems to depend on the structure of the stimulus presented first. Experiment III investigates perceptual interactions between implosive and explosive transitions by preceding stimuli from a /bɛ/-/dɛ/ continuum with either /ab/ or /ad/, or following stimuli from an /ab/-/ad/ continuum with either /bɛ/ or /dɛ/. Precursor/postcursor effects on the stimuli from the acoustic continua are measured on a six-point rating scale. At a closure duration of 25 msec, the implosive transitions exert a pronounced assimilative effect on the perception of the explosive transitions, although the former are not perceived as a separate phonemic event. At a closure duration of 265 msec,

---

Acknowledgment: These experiments were supported by NICHD Grant HD01994 to the Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

explosive transitions exert a slight contrastive effect on implosive transitions (now perceived as a separate consonant) but not vice versa.

The present experiments continue and extend research reported earlier by Dorman, Raphael, Liberman, and Repp (1975) and Repp (1975, 1976a, 1976b). For a general introduction, the reader is referred to these earlier articles.

EXPERIMENT I

This experiment had two parts: an identification task and a discrimination task. Dorman et al. (1975) demonstrated that a VC_1-C_2V utterance, for example, /ɛb-dɛ/, tends to be perceived as VC_2V (that is, /ɛdɛ/) if the stop closure interval is artificially shortened. A period of 50-80 msec of silence is needed to identify C_1 correctly. In an experiment using synthetic stimuli similar to those used in the present studies, Becky Treiman[1] found asymptotic identification performance at a closure period of 60 msec. Informal observations of my own showed that it is not absolutely necessary to follow the implosive transitions (C_1) with conflicting explosive transitions (C_2) for the perception of C_1 to be impaired; similar interference also seemed to occur in VC-V utterances, that is, when the implosive transitions were followed (after a short period of silence) by a steady-state vowel that did not provide conflicting information about the place of articulation of the stop consonant. This effect was to be demonstrated more formally in Task 1 of the present experiment.

The results of Dorman et al. and Treiman were obtained in identification tasks, where the subjects simply wrote down what they heard. While /ɛb-dɛ/ with a very short closure period may sound like /ɛdɛ/, the question remains whether it sounds exactly like a "real" /ɛdɛ/ (that is, /ɛd-dɛ/) with the same closure period. Although the two utterances are phonetically alike, they may still have a discriminably different auditory quality, or one may be a less convincing instance of /ɛdɛ/ than the other. Recently, I demonstrated (Repp, 1976b) that it is very difficult to discriminate /ɛd-dɛ/ from /ɛ-dɛ/, where the implosive transitions have been substituted with steady-state vowel formants. Performance in this task approached chance at a closure period of 65 msec, the shortest interval used in this earlier study. To explore both issues further, three types of utterances were tested for their discriminability from each other in Task 2 of the present experiment. The three stimulus types were VC-CV (for example, /ab-dɛ/, heard as /adɛ/ at short closure intervals), V(C)-CV (for example, /ad-dɛ/, which is always heard as /adɛ/ at the closure durations used here), and V-CV (for example, /a-dɛ/, which is heard as /adɛ/ at short closure durations and as /a-dɛ/ -- with a perceptible pause between initial vowel and consonant -- at longer closure durations). These stimuli differed only in the portions immediately preceding the silent closure interval: the implosive transitions were either incompatible with the explosive transitions (VC-CV), compatible (V(C)-CV), or completely absent (V-CV). Task 2 of Experiment I was designed to determine

---

[1]This experiment was conducted by Ms. Treiman, with my assistance, to fulfill a course requirement at Yale University. No formal write-up is available.

the functions that relate discrimination accuracy to closure duration for the three pairs of stimulus types.

Method

Subjects. The subjects were 10 relatively inexperienced listeners and myself.[2] Eight of the subjects had previously participated in Experiment II (described later in this paper) and thus had been exposed to stimuli very similar to those in the present experiment.

Stimuli. The stimuli were derived from those used in my earlier studies (see Repp, 1976b, for details). The basic stimuli were /abɛ/, /abi/, /adɛ/, and /adi/, synthesized on the Haskins Laboratories parallel formant synthesizer. In the stimuli for Task 1, the explosive transitions of the medial stop consonant were replaced with the steady-state formants of the following vowel. This resulted in /ab-ɛ/, /ab-i/, /ad-ɛ/, and /ad-i/, that is, /ab/ and /ad/ followed by either /ɛ/or /i/ after a variable closure duration. The closure intervals ranged from 0 to 125 msec in 25-msec steps. The resulting 24 stimuli were recorded in five different randomizations with interstimulus intervals (ISIs) of 3 sec.[3] This series was preceded by a random sequence of 10 /ab/ and 10 /ad/ syllables.

Three types of stimuli were prepared for Task 2. The original stimuli represented the V(C)-CV set in which both implosive and explosive transitions were appropriate for the same place of articulation. VC-CV utterances /ab-dɛ/, /ab-di/, /ad-bɛ/, and /ad-bi/ were obtained by interchanging the VC portions of the V(C)-CV stimuli. V-CV stimuli /a-bɛ/, /a-bi/, /a-dɛ/, and /a-di/ were obtained by replacing the implosive transitions with the steady-state formants of the initial vowel, holding formant amplitudes constant. The closure durations used ranged from 0 to 100 msec in 25-msec steps, except for the practice trials, where the stimuli had a 250-msec closure interval.

There were three discrimination conditions: V(C)-CV vs. VC-CV, V(C)-CV vs. V-CV, and VC-CV vs. V-CV. An AXB paradigm was used, that is, the first and the last stimulus in a triad were always different from each other, and the second stimulus was identical with either the first or the third. The stimuli in a triad always had the same closure duration and the same CV portions; they differed only in the information immediately preceding the closure interval. In each condition, there were 80 AXB triads, resulting from 4 stimuli with 5 closure durations in 4 AXB configurations (AAB, ABB, BAA, BBA). Each series was randomized and recorded as a separate block, preceded by 16 practice trials (stimuli with 250-msec closure intervals).

---

[2] My own data were included because they were not qualitatively different from those of the other subjects (although I made fewer errors) and because they had also been included in Experiment II of Repp (1976b), which was to be compared with the present results.

[3] Note that, in this paper, the term ISI never denotes the brief silent interval between the VC and CV (or V) portions of stimuli, which is always referred to as closure interval.

The within-triad ISI was 1 sec; the between-triad ISI, 3 sec.

Procedure. All subjects first did the identification task. After listening to the short practice list of VC syllables, the subjects listened twice to the VC-V identification series. The first time they were instructed to write down B when they heard /ab-ɛ/ or /ab-i/, D when they heard /ad-ɛ/ or /ad-i/, and 0 when they heard no consonant at all, that is, /a-ɛ/ or /a-i/. In the second run, 0-responses were no longer permitted, and a forced choice between B and D had to be made for each stimulus.

The sequence of the three discrimination conditions was approximately counterbalanced across subjects. The structure of the stimuli was explained before each condition, so that the subjects knew quite well what they were listening to and what they were trying to discriminate. The responses were A and B, whichever the X stimulus in the AXB triad equalled.

Other procedural details were the same as in previous studies (see Repp, 1976b).

Results

Task 1: VC-V Identification. All subjects identified the 20 practice VC syllables without difficulty. Only a single error was committed. The results of the first run through the VC-V identification task are shown in Figure 1a. The dashed line represents 0-responses, the solid line the total error rate (that is, 0-responses plus confusion errors). The difference between the two functions is the percentage of confusions, which did not change at all with closure duration. The percentage of 0-responses declined rapidly over the first 25 msec and then more slowly.

Estimates of performance level in Run 1 may be obtained by assuming that, if forced to guess instead of responding 0, the subjects would have been correct 50 percent of the time. These estimates are shown in Figure 1b together with the results of the second run (where 0-responses were not permitted). It can be seen that performance was close to chance when there was no closure interval at all, but it improved rapidly as closure duration increased. An asymptote seemed to be reached at a closure duration of 100 msec; however, note that the asymptotic error rate was much higher than for VC syllables in isolation! Performance in Run 2 was better than in Run 1. This may reflect not only practice effects, but also the incorrectness of the assumption that all 0-responses were equivalent to random guesses.

A 4-way analysis of variance was performed on the data in Figure 1b, with consonant and (final) vowel as additional factors. The effects of runs ($F_{1,10} = 26.6$, $p < .01$) and of closure duration ($F_{5,50} = 33.5$, $p << .01$) were highly significant. In addition, however, there was a significant effect of consonant ($F_{1,10} = 5.5$, $p < .05$) and a highly significant consonant x vowel interaction ($F_{1,10} = 17.2$, $p < .01$). This interaction is shown in Figure 2. It is evident that, when followed by a vowel, /ab/ was much easier to identify than /ad/, especially at longer closure durations, where /ad/ stimuli were solely responsible for the high error rates. In addition, /ad-i/ was much more difficult than /ad-ɛ/, but /ab-ɛ/ was more difficult than /ab-i/.

FIGURE 1

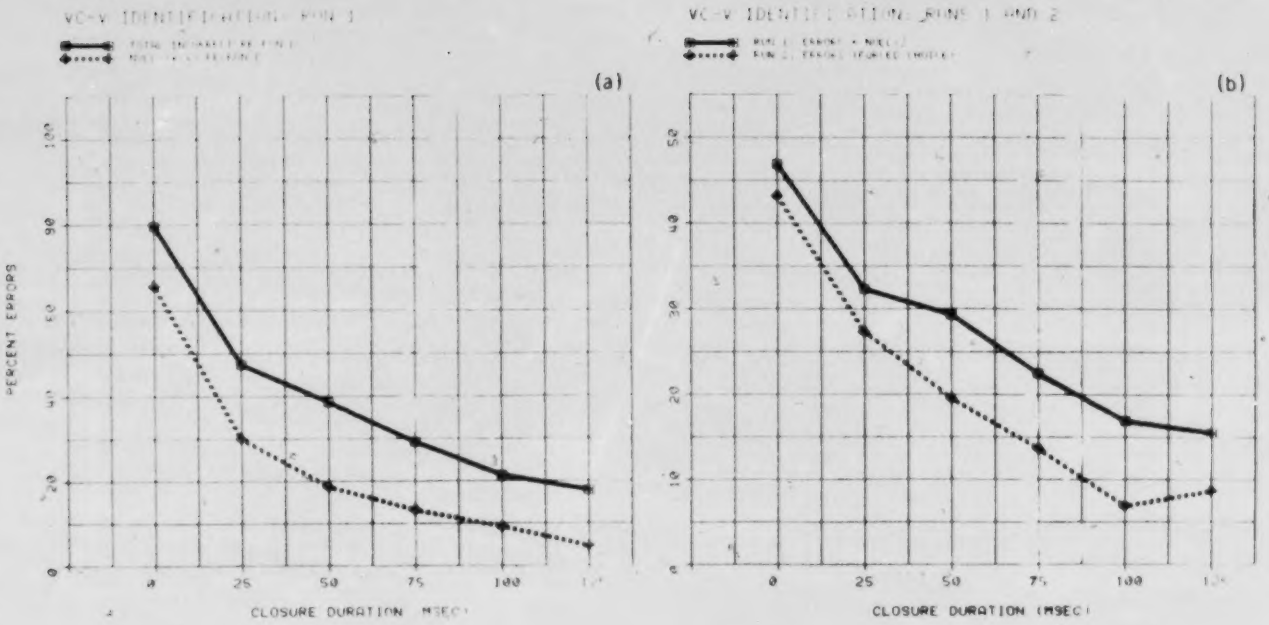


Figure 1: VC-V identification errors as a function of closure duration.

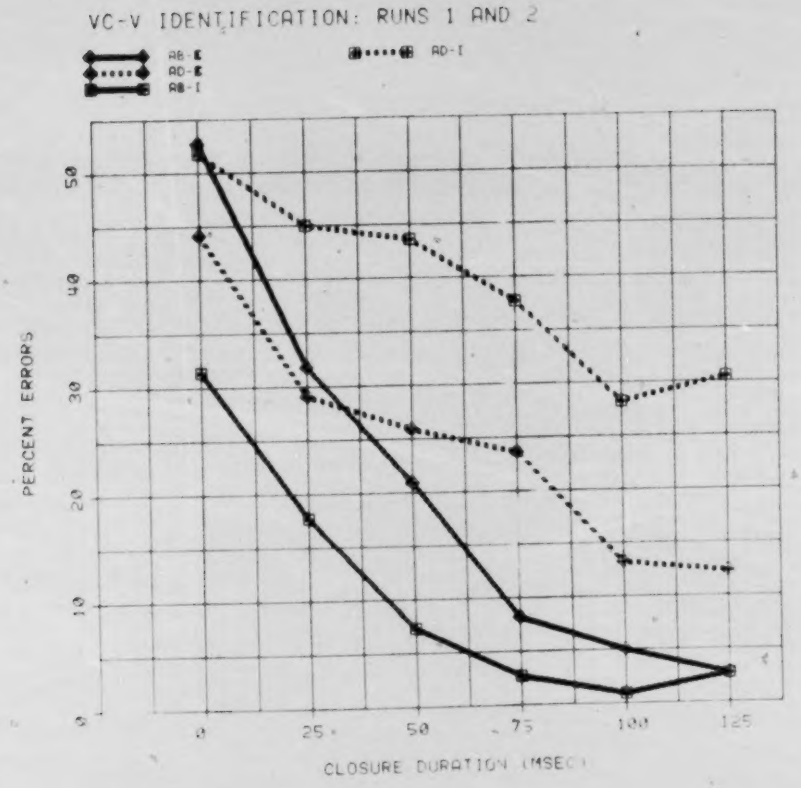


Figure 2: Differences between consonants, and effects of final vowels in VC-V identification.

Task 2: VCV Discrimination. As expected, the V(C)-CV vs. V-CV (compatible vs. absent implosive transitions) condition was the most difficult. This was evident already in the practice trials, where the average error rate was 16.5 percent in this condition, but only 6.8 and 6.3 percent, respectively, in the other two conditions. The results for the shorter closure durations are shown in Figure 3. The compatible-absent condition was more difficult than the incompatible-absent condition, which in turn was more difficult than the compatible-incompatible condition. Error rates declined steadily as closure duration increased, but were still considerably above the practice trial error rate at the longest closure duration, which thus does not represent the asymptote.

A 4-way analysis of variance showed not only highly significant effects of condition ($F_{2,20} = 17.3$, $p < .01$) and closure duration ($F_{4,40} = 38.1$, $p << .01$), but also a significant effect of vowel ($F_{1,10} = 7.9$, $p < .05$) and a significant condition x consonant interaction ($F_{2,20} = 13.6$, $p < .01$). Since none of these effects interacted with closure duration, the data were collapsed over this factor and each condition was analyzed separately in a 2-way analysis of variance.

In the compatible-absent condition, there was only a significant effect of consonant ($F_{1,10} = 13.1$, $p < .01$) which is shown in Figure 4a. Quite obviously, the presence of implosive transitions was much more difficult to detect in /ab/ (B) than in /ad/ (D). For /ab/, performance remained at chance level up to a closure duration of 50 msec or more, while for /ad/ the error percentage decreased almost linearly from the beginning. These resuls are in excellent agreement with those of Repp (1976b, Experiment II, Task 3, Figure 5) where exactly the same difference was found at slightly longer closure durations.

In the compatible-incompatible condition, there was only a marginally significant effect of vowel ($F_{1,10} = 5.5$, $p < .05$). Stimuli ending in /-ɛ/ were easier than stimuli ending in /-i/, but this difference was present only at two closure durations (25 and 75 msec). A similar effect was present in the incompatible-absent condition but did not quite reach significance ($F_{1,10} = 4.2$, $p < .10$). However, in the latter condition, there was a significant effect of consonant ($F_{1,10} = 8.9$, $p < .02$). As indicated by the significant condition x consonant interaction obtained earlier, this effect was in the opposite direction of that in the compatible-absent condition. However, since the consonant factor in this earlier analysis reflected the nature of the explosive transitions (which alone were constant from condition to condition), it is obvious that the consonant effect in the incompatible-absent condition (shown in Figure 4b) was the same as that in the compatible-absent condition in terms of implosive transitions. Thus, the presence of implosive labial transitions was more difficult to detect, regardless of whether they were followed by compatible or incompatible explosive transitions. That the two consonant effects shown in Figure 4 were in fact due to the implosive transitions and not the explosive transitions, is also supported by the complete absence of a consonant effect in the compatible-incompatible condition, where the consonant factor reflected only variation in the explosive transitions.

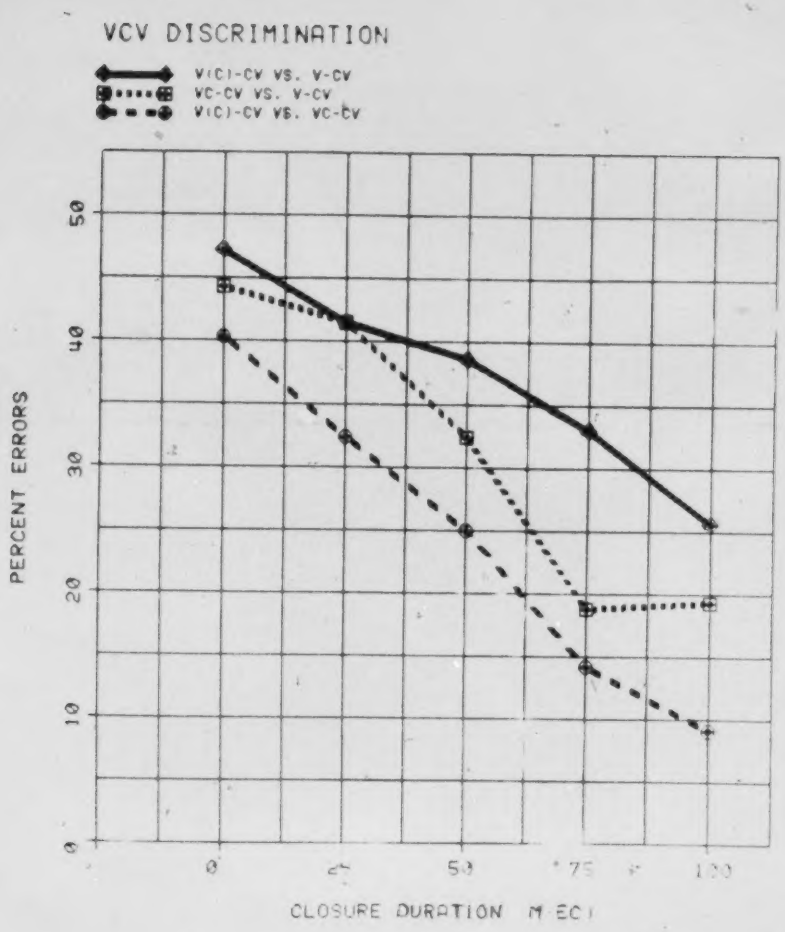


Figure 3: Error percentages as a function of closure duration in three different discrimination tasks.

FIGURE 4

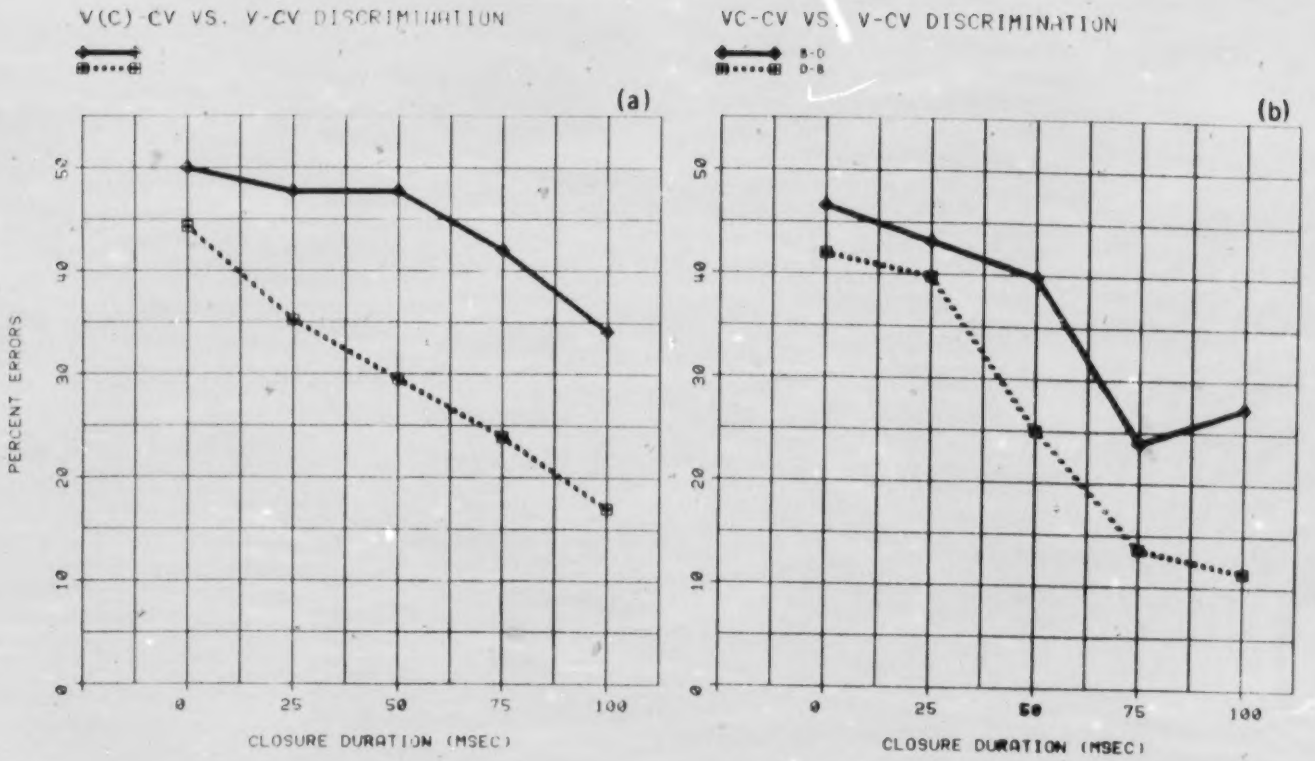


Figure 4: Differences between consonants (implosive transitions) in two discrimination tasks.

Discussion

The VC-V identification task demonstrates that it is not necessary to follow implosive transitions with explosive transitions to produce interference at short closure durations. A steady-state vowel is sufficient. No direct comparison of the two effects was conducted here, but comparison with the results of Dorman et al. (1975) and Treiman (see Footnote 1) suggests that the interference of a steady-state vowel is somewhat less severe than that of incompatible explosive transitions at short closure durations, but that it extends to longer closure durations. Overall, the two effects do not seem fundamentally different from each other.[4]

This finding is compatible with both explanations that have been forwarded for the interference effect. One explanation claims it is true recognition backward masking, that is, interruption of processing (Massaro, 1975); the other refers to links between perception and production and assumes that the perceptual system refuses to deal with speechlike sounds that are impossible to articulate (Dorman et al., 1975; Liberman, 1975). It is certainly true that a VC-V utterance with a perfectly steady-state final vowel could not be pronounced with closure durations as short as the ones used here. There is a more specific implication for the backward-masking hypothesis: if it is correct, then the interruption of processing of the implosive transitions probably does not take place in a mechanism specialized for the perception of stop consonants or place of articulation, since the masking vowel presumably does not engage this processing mechanism. Rather, we seem to be dealing with a more general auditory interference with the perception of implosive transitions.

The large differences due to consonants (/b/ vs. /d/) and vowels (/ɛ/ vs. /i/) were quite surprising. On the basis of the acoustic characteristics of the target stimuli alone, a consonant effect in the opposite direction might have been expected. /ab/ differed from /ad/ not only in the second- and third-formant transitions, but it also had a shorter first-formant (F_1) transition with a higher terminal frequency.[5] Since the F_1 transition is an important manner cue, one might have expected /b/ to be less "stop-like" and therefore more susceptible to interference than /d/. Instead, /d/ was much more affected by the following vowel. Closer inspection of the data from Run 1 showed, however, that this difference was primarily due to genuine misidentifications of /d/ as /b/; there was a much smaller difference between the two consonants in terms of 0-responses, which reflect detection of the manner cue. Thus, it was alveolar place of articulation that specifically suffered from interference. Moreover, as

---

[4]Malmberg (1955) noted long ago that the consonant of VC-V stimuli is perceptually grouped with the final vowel when the closure duration is made very short; he did not mention any interference effect. This difference may be ascribed to Malmberg's use of sophisticated listeners, and perhaps to the identity of the initial and final vowels in his stimuli.

[5]This difference was not really intended but somehow crept into the original stimulus set and then was carried along. It was also present in the earlier experiments (Repp, 1976a, 1976b) and was eliminated only in the present Experiments II and III.

Figure 1a shows, there was virtually no decline in the frequency of genuine confusions as closure duration increased. Even at a closure duration of 125 msec, a large proportion of /d/ stimuli was labeled /b/, despite the fact that the VC portions were perfectly identifiable in isolation. This was surprising since the final /d/ was, in a sense, acoustically more "prominent" than the final /b/, due to its steeply rising third-formant transitions (cf. also Repp, 1976b). Precisely for this reason, however, /ad/ perhaps sounded less natural than /ab/, and this may have been responsible for its poor identification when followed by a steady-state vowel.

The interaction between consonants and vowels (Figure 2) is even more intriguing. Closer inspection of the data from Run 1 indicates that the differential effect of the two vowel masks on /ab/ was entirely due to 0-responses, while, with /ad/, both 0-responses and genuine errors showed a large vowel effect. The differential effect of the two vowels on the detectability of the manner cue (implosive F_1 transitions) may have been due to their different F_1 frequencies. However, why did it interact with the final consonants? As mentioned earlier, /ad/ had a longer F_1 transition with a lower offset frequency than /ab/; /i/ had a lower F_1 than /ɛ/. Thus, the F_1 of /i/ (279 Hz) was closer to the F_1 offset of /ad/ (381 Hz), and the F_1 of /ɛ/ (535 Hz) was closer to the F_1 offset of /ab/ (560 Hz). This relative continuity of F_1 may have led to the perceptual illusion of a transition between two vowels without any intervening silence. We thus arrive at the -- admittedly speculative -- hypothesis that a listener will be less likely to perceive an implosive F_1 transition as a stop manner cue if it points towards the F_1 frequency of a following vowel.

The devastating effect of /i/ on the perception of /ad/ remains to be explained. Perhaps the relative continuity of the second and third formants (F_2 and F_3) can provide an explanation. /ad/ had rising implosive F_2 and F_3 transitions; the F_2 offset (1459 Hz) was below the F_2 of /ɛ/ (1840 Hz) and far below the F_2 of /i/ (2298 Hz), while the F_3 offset (3363 Hz) was above the F_3 of /i/ (3029 Hz) and far above the F_3 of /ɛ/ (2527 Hz). A formant continuity interpretation would be possible only for F_3 but not for F_2, for which the relationships are reversed. However, to the degree that the F_3 transition was responsible for the somewhat artificial sound of /ad/, its relative continuity with the F_3 of the following vowel (especially /i/) may have specifically harmed /d/ identification.

Thus, the results point towards frequency-specific interactions between the implosive transitions and the following vowel. Relative continuity of formants across the intervening silence seems to make it more difficult to perceive manner and place conveyed by implosive transitions. This effect is reasonable from an auditory information processing viewpoint. However, an interpretation in terms of articulatory relationships may also be possible, since auditory and articulatory variables are highly correlated.

Turning to the results of the discrimination task, note first that the general interfering effect of C_2 on C_1 in VC-CV utterances was confirmed. As closure duration was decreased, VC-CV became increasingly more difficult to discriminate from V(C)-CV and V-CV. The extent and time course of the effect were not only similar to those reported earlier by Dorman et al. (1975), but they also paralleled the results in the VC-V identification task, confirming

that backward interference by a steady-state vowel and by an incompatible CV syllable are basically similar effects. The fact that the present "masking" functions extend over a wider range of closure durations than those of Dorman et al. (1975) may be due to differences in stimulus structure and methodology.

Despite the fact that V(C)-CV and V-CV utterances were very difficult to discriminate from each other, VC-CV stimuli were consistently easier to discriminate from V(C)-CV stimuli than from V-CV stimuli. In the V(C)-CV vs. VC-CV condition, the difference consisted in a large difference in F_2 and F_3 and a small difference in F_1. In the V-CV vs. VC-CV condition, the difference in F_1 was larger, but that in F_2 and F_3 was smaller. Apparently, then, it was the difference in the higher formants that was more important for discrimination performance. The acoustic differences to be discriminated in VC-CV vs. V-CV and in V(C)-CV vs. V-CV were about equivalent, and indeed performance in the two conditions was similar at the two shortest closure durations (see Figure 4). At longer closure durations, the former condition had an advantage as the difference in phonetic structure began to emerge. Thus, the results suggest that discriminations at very short closure durations were made primarily on the basis of auditory differences (very inefficiently), while at longer closure durations, phonetic distinctions played an increasing role. The difference between VC-CV vs. V(C)-CV and VC-CV vs. V-CV at longer closure durations may also reflect a phonetic factor, as suggested by my own observations: when V-CV stimuli were paired with VC-CV stimuli, the VC-CV context sometimes induced the V-CV stimuli to be heard as VC-CV too, thus reducing discrimination accuracy. In V(C)-CV stimuli, the presence of (compatible) implosive transitions apparently prevented such phonetic illusions. It also seems that they did not occur in V(C)-CV vs. V-CV discrimination, so that the better-than-chance discriminability of these stimuli must be ascribed to an auditory cue -- a slight discontinuity between initial vowel and consonant in VC-V stimuli that became noticeable as closure duration increased.

The strong consonant effect in the V(C)-CV vs. V-CV condition replicated the effect found by Repp (1976b). Most likely, it was due to the perceptible acoustic difference between the implosive transitions of the two consonants. The steeper F_1 and F_3 transitions of /ad/ and its resulting somewhat strident sound insured its better discriminability from the steady-state vowel. The consonant effect was less pronounced in the VC-CV vs. V-CV condition, perhaps because of the higher performance level there, which resulted from the additional phonetic factor aiding discrimination.

While both the above conditions involved the discrimination of a VC stimulus from a V stimulus in the presence of a constant CV "mask" -- which amounts to detecting the presence of implosive transitions -- the third condition, V(C)-CV vs. VC-CV, involved the discrimination of two different types of implosive transitions. Thus, unlike the other two conditions, exactly the same target discrimination was required on every trial, and only the CV mask varied. In contrast to the variation in implosive transitions, the variation in explosive transitions had no effect on performance.

EXPERIMENT II

This experiment was a follow-up to Experiment I of Repp (1976b) and very similar in design. Repp (1975, 1976a) showed that "same" judgments about the medial consonants of two successive VCV (that is, V(C)-CV) utterances have shorter latencies when the final vowels are the same than when they are different, and that this effect persists when the closure period is increased. Since the absolute latencies also increased with closure duration, it seemed that the subjects based their decisions solely on the CV portions of the stimuli. Repp (1976b, Experiment I) used a design that randomly mixed VC and VCV utterances, in order to force the listeners to focus on the implosive transitions. This procedure was successful in so far as the latencies no longer increased systematically with the closure duration of the VCV stimuli. Paradoxically, however, the effect of the final vowel on "same" latencies did not disappear at long closure durations -- a result that could not be explained, since the latencies seemed to indicate that the subjects relied on the implosive transitions alone, which were independent of the final vowel.

Repp (1976b, Experiment II) employed a simpler choice-reaction time task to get at the same problem. By presenting VCV stimuli with and without implosive transitions (that is, V(C)-CV and V-CV stimuli) and varying closure duration, I demonstrated that response latencies for deciding whether a stimulus began with /ab/ or /ad/ increased with closure duration for V-CV stimuli, but not for V(C)-CV stimuli. Clearly, then, the listeners were paying selective attention to the VC portion of the stimuli. However, latencies for isolated VC stimuli were faster than for V(C)-CV stimuli, which showed that the following CV portion in V(C)-CV stimuli still affected the decision process.

The alternative, and perhaps more obvious, procedure to investigate the influence of the CV portion on decisions about the VC portion is to remove the explosive transitions and compare latencies for V(C)-CV and VC-V stimuli. This approach was taken in the present experiment, after some hesitation. While removing the implosive transitions of a V(C)-CV stimulus has little perceptual consequence (V(C)-CV and V-CV stimuli sound extremely similar at short closure durations -- cf. Experiment I), removing the explosive transitions has a much more disturbing effect: V(C)-CV and VC-V stimuli sound differently, especially at short closure durations, where the consonant in VC-V stimuli is difficult to perceive (cf. Experiment I). Thus, high error rates were to be expected, but I nevertheless found the experiment worth attempting.

The present experiment consisted of three tasks. Task 1 served to familiarize the listener with the basic target stimuli; it required a simple forced-choice classification of the two standard VC syllables, /ab/ and /ad/. Task 2 was also a consonant classification task, but here most of the VC targets were followed by either a phonetically compatible CV syllable or by a steady-state vowel, after one of two closure intervals. It was expected that whatever influence the explosive transitions exerted on consonant judgments would be absent in VC-V stimuli, so that latencies were expected to be faster for VC-V stimuli than for V(C)-CV stimuli. However, since the intelligibility of the VC-V consonant suffered at short closure durations, it was

considered possible that the faster latencies for VC-V stimuli would emerge only at the longer closure duration.

Task 3 was a same-different reaction time (RT) task. Here, as in the previous experiments (Repp, 1975, 1976a, 1976b -- Experiment I, Task 3), the effect of principal interest was the influence of the final vowel on the latency of "same" judgments. The design included V(C)-CV and VC-V stimuli with two closure durations, as well as VC stimuli, in various combinations with each other. Repp (1976b) hypothesized that, in V(C)-CV pairs, the subjects compared the explosive transitions instead of the implosive transitions on some trials, leading to an effect of the final vowel even at long closure durations. If this interpretation is correct, the final-vowel effect should disappear in VC-V pairs that do not contain any explosive transitions. On the other hand, if the effect of the final vowel is due to some more general perceptual integration, it should be present in VC-V stimulus pairs as well (perhaps in reduced magnitude). Again, some effect at short closure durations was to be expected simply because of the interfering effect of the final vowel; the more interesting condition was the long closure duration.

Although these hypotheses were formulated in terms of latencies, the experiment contained a safeguard against the possibility that RTs would show too much variability due to the relative difficulty of the task for inexperienced listeners. Earlier experiments have shown that error rates are highly correlated with latencies in this type of task, and as task difficulty increases, they become a more reliable dependent variable than the latencies themselves. Most of the hypotheses could therefore be replaced by substituting "fewer errors" for "faster latencies". As it turned out, I had to rely heavily on the error rates in interpreting the results of the present experiment.

Method

Subjects. Ten volunteer subjects participated, all of them relatively inexperienced in this type of experiment.

Stimuli. The same basic set of V(C)-CV stimuli was used as in the earlier experiments (/abɛ/, /abi/, /adɛ/, /adi/). VC-V stimuli were generated by replacing the explosive transitions with steady-state vowel formants, as in Task 1 of Experiment I. Closure durations were 100 and 250 msec. VC stimuli consisted only of the stimulus portions preceding the silent closure interval. One slight difference between the present stimuli and those of earlier experiments was that the F_1 transitions of /ab/, originally shorter than those of /ad/, were made equally long. While this may have increased the detectability of implosive labial transitions in V(C)-CV stimuli (cf. Figure 4a), it hardly affected the intelligibility of VC-V stimuli in which, primarily, /ad/ suffered from the following vowel (cf. Figure 2).

The initial VC list (Task 1) contained 50 stimuli in random order with ISIs of 3,555 msec. The choice-RT sequence (Task 2) contained 100 stimuli presented in five individually randomized blocks of 20. Each block contained 16 VCV stimuli (four basic stimuli with or without explosive transitions at two closure durations) and 4 VC stimuli. The ISI covaried with closure duration and stimulus type; it was the stimulus onset (or VC offset)

asynchrony that was held constant at 3,740 msec. The tape for Task 3 contained two individually randomized blocks of 144 stimulus pairs. Each block contained all pairwise combinations of the four V(C)-CV stimuli and all pairwise combinations of the four VC-V stimuli at each of the two closure durations, resulting in 2 x 2 x 16 = 64 stimulus pairs; plus all combinations of the two VC stimuli with all V(C)-CV and VC-V stimuli at each closure duration, resulting in another 2 x 2 x 16 = 64 stimulus pairs; plus four replications of the four VC combinations. Note that the two stimuli in a VCV pair always were of the same type (V(C)-CV or VC-V) and had the same closure duration. The within-pair onset asynchrony was constant at 1 sec; the between-pair onset asynchrony (from the onset of the second stimulus in a pair to the onset of the first stimulus of the next pair) was 3,740 msec.

Procedure. Equipment, procedure, and analysis were almost exactly identical to those of Repp (1976b, Experiment I). Only the main features shall be repeated here. In Tasks 1 and 2, the subjects pressed one response key for /ab/ and the other for /ad/, ignoring the final vowel, if present. The response-hand assignment was varied from subject to subject. In Task 3, all subjects responded "same" with the (preferred) right hand and "different" with the left. It was emphasized to respond as quickly as possible, to ignore the final vowels, and not to wait for the end of an utterance before responding. It was mentioned that some stimuli might be a little more difficult to identify than others. Subjects were asked to "correct" their own errors (if realized) by quickly pressing the other key. (This procedure was found useful in earlier studies but had been neglected in the earlier experiments of this series.) Each subject listened to the two blocks twice, that is, to 4 x 144 = 576 stimulus pairs altogether. All tasks were preceded by a few minutes of practice selected randomly from the tapes.

Data analysis was conducted on the median RTs of correct responses calculated from 25 stimulus replications in Task 1, from 5 replications (10 for VC stimuli) in Task 2, and from 8 responses (16 for VC pairs) in Task 3. These eight responses in Task 3 resulted from cross-classifying the responses according to the factors blocks (1 and 2 vs. 3 and 4), stimulus types (V(C)-CV vs. VC-V), closure duration (100 vs. 250 msec), same/different consonant, and same/different vowel, which left eight responses per cell. Pairs containing VC stimuli were analyzed separately from the other (structurally homogeneous) pairs; the factorial design was similar, except that temporal order (VC first or second) replaced the same/different vowel factor. VC pairs were not included in this analysis.

RTs were measured from VC offset in each case. Errors corrected by the subjects themselves were omitted from analysis, since earlier studies had indicated that they were mostly due to response anticipations or response hand confusions and not related to the experimental conditions. Except for individual differences in frequency, they showed no obvious pattern in the present experiment either.

Results

Task 1: VC Classification. The subjects had little difficulty in classifying the VC syllables in isolation. The overall error rate was 1.8 percent, excluding corrected errors (2.0 percent). They consisted of 8

errors with /ad/ (3.2 percent) and only 1 error with /ab/ (0.4 percent). RTs were faster to /ab/ (368 msec) than to /ad/ (410 msec). This difference was shown by eight of the ten subjects and was significant ($F_{1,10} = 13.21$, $p < .01$). It is in the opposite direction of the difference found by Repp (1976b -- Experiment I, Task 1). In fact, while the average RTs to /ad/ are comparable in the two studies, those to /ab/ were faster in the present study by over 100 msec. This difference most likely reflects the change in the F_1 transition of /ab/.

Task 2: Choice-RT Task. The results of the choice-RT task are shown in Figure 5. Figure 5a shows the latencies, Figure 5b the error rates. Both figures show an interaction between stimulus type and closure duration. While closure duration had relatively little effect in V(C)-CV stimuli, performance with VC-V stimuli was much better at the long closure duration than at the short one. This was expected, because the final vowel interfered with the perception of the implosive transitions at the 100-msec closure duration; the error rate was correspondingly high. It is interesting, however, that at the 100-msec closure duration, VC-V RTs were hardly longer than V(C)-CV RTs, despite the large difference in error rates, and at the 250-msec closure duration, VC-V RTs were actually faster than V(C)-CV RTs, although VC-V stimuli continued to exhibit a slightly higher error rate. Thus, although error rates and latencies tend to be positively correlated, sometimes one measure shows a difference where the other does not (cf. Repp, 1976b, for similar observations).

Unfortunately, the RT effects did not reach significance due to large individual differences and high variability. A 4-way analysis of variance (stimulus types, closure durations, consonants, vowels) yielded no significant effects. Transformations of the data or eliminating subjects with exceptionally long RTs did not help. Thus, no firm conclusions can be drawn from the RT pattern in Figure 5a.

The error patterns were more consistent, although the majority of the errors was contributed by a few subjects. The overall error rate was 9.5 percent, excluding corrected errors (3.5 percent). In addition to the effects of stimulus type and closure duration evident in Figure 5b, there were the expected large differences between individual stimuli: /adi/ (26.0 percent), /adɛ/ (10.5 percent), /abɛ/ (3.0 percent), /abi/ (2.0 percent). Thus, the large majority of the errors consisted in alveolar-to-labial confusions. For VC-V stimuli with a closure duration of 100 msec, the error rates for the four individual stimuli were 42.0, 34.0, 6.0, and 10.0, respectively -- considerably higher than in Experiment I, Task 1 (Figure 2). This difference probably reflects the more stringent demands of the present task and perhaps context effects; however, the pattern agrees with the results shown in Figure 2.

Error rates for VC stimuli were comparable to those for other stimuli at the longer closure duration (Figure 5b). However, RTs tended to be faster for VC stimuli than for VCV stimuli (Figure 5b). VC stimuli in Task 2 exhibited both higher error rates and slower RTs than the VC stimuli in Task 1 -- a context effect also obtained by Repp (1976b).

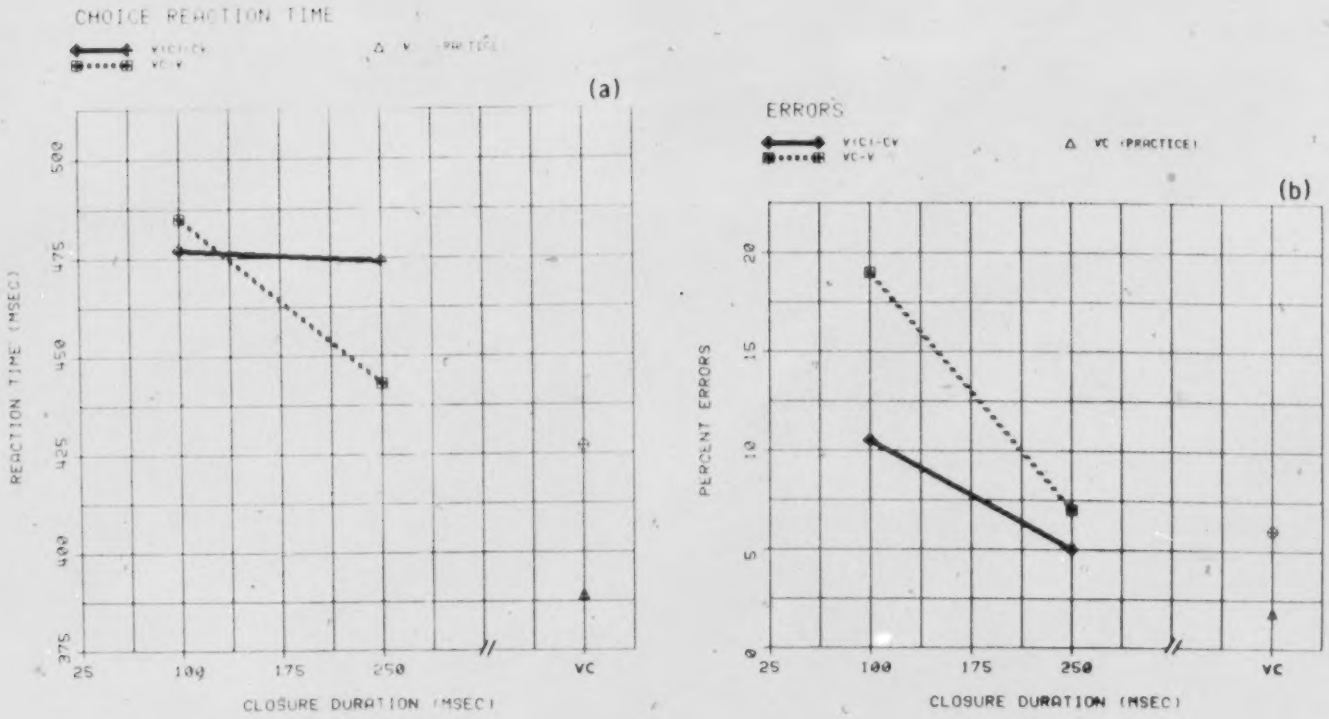


FIGURE 5

Figure 5: Mean latencies of correct responses and error rates in the choice-RT task as a function of closure duration.

Task 3: Same-Different Reaction Time. The results of Task 3 are shown in Figures 6 and 7. Figure 6 shows the data for VCV pairs (that is, pairs not containing VC stimuli). It has four panels: panels a and b (above) show RTs, panels b and c (below) the corresponding error rates. Panels a and c (on the left) are for V(C)-CV stimuli, panels b and d (on the right) are for VC-V stimuli.

The latency data were analyzed in a 5-way analysis of variance that yielded several significant effects. However, the effects that were not significant provided equally interesting information: there was no significant practice (block) effect, no significant overall increase in RTs with closure duration, no significant overall difference between V(C)-CV and VC-V pairs, and (surprisingly) no significant difference between "same" and "different" RTs (that were confounded with right vs. left response hand). The only main effect that reached significance was that of same/different vowel ($F_{1,9} = 8.43$, $p < .05$), with faster overall latencies when vowels were the same. Several higher-order interactions reached significance but do not merit extensive discussion. They were primarily due to the precipitous decline in VC-V "different" latencies with closure duration where the words were the same (cf. Figure 6b).

It is evident from Figures 6a and 6b that both stimulus types showed an effect of the final vowel on "same" latencies. The effect was in the expected direction (faster RTs when the vowels were the same) and did not decrease as closure duration increased. The effect of the final vowel on "different" latencies was not consistent, on the other hand, and seemed to interact with stimulus types as well as closure duration. The result that the final vowels had a consistent effect on "same" responses only is in agreement with earlier experiments, and so is the absence of a decline of this effect as closure duration increased.

To clarify the statistical reliability of the effect, a separate analysis of variance was conducted on "same" latencies only. The main effect of same/different vowel reached significance ($F_{1,9} = 5.13$, $p < .05$) and did not interact with any other factor. The only other significant effect was an uninterpretable 3-way interaction between the other three factors (blocks, closure durations, stimulus types).

Because of the high error rates and the great variability of the latencies, the error pattern was likely to provide a more direct and consistent indicator of the major experimental effects. Figures 6c and 6d show quite clearly that (1) more errors were made on "different" trials than on "same" trials (that is, incorrect "same" responses were more frequent than incorrect "different" responses), (2) "different" trials had much higher error rates with VC-V pairs than with V(C)-CV pairs, (3) "different" errors (that is, incorrect "same" responses) decreased as closure duration increased, but "same" errors remained roughly constant, and (4) the same/different vowel factor had a clear effect only on "same" errors and was independent of closure duration and stimulus type. Error and latency patterns for "same" responses are in good agreement, which in part reflects the greater reliability of "same" latencies because of the lower error rates on "same" trials. There was no increase in accuracy over blocks. All effects just mentioned were highly significant in an analysis of variance,

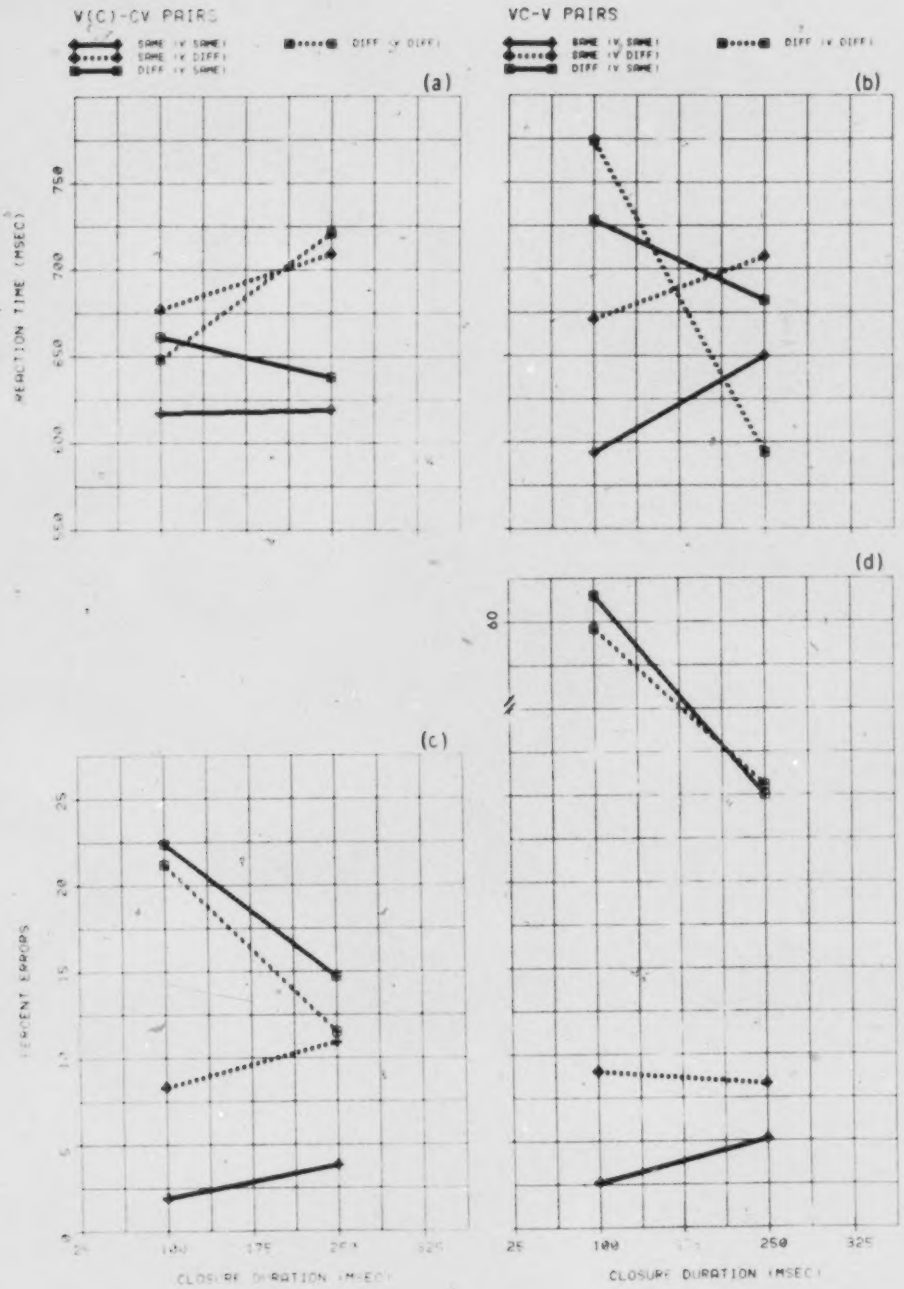


Figure 6: Mean latencies of correct responses and error rates in the same-different task: V(C)-CV pairs and VC-V pairs.

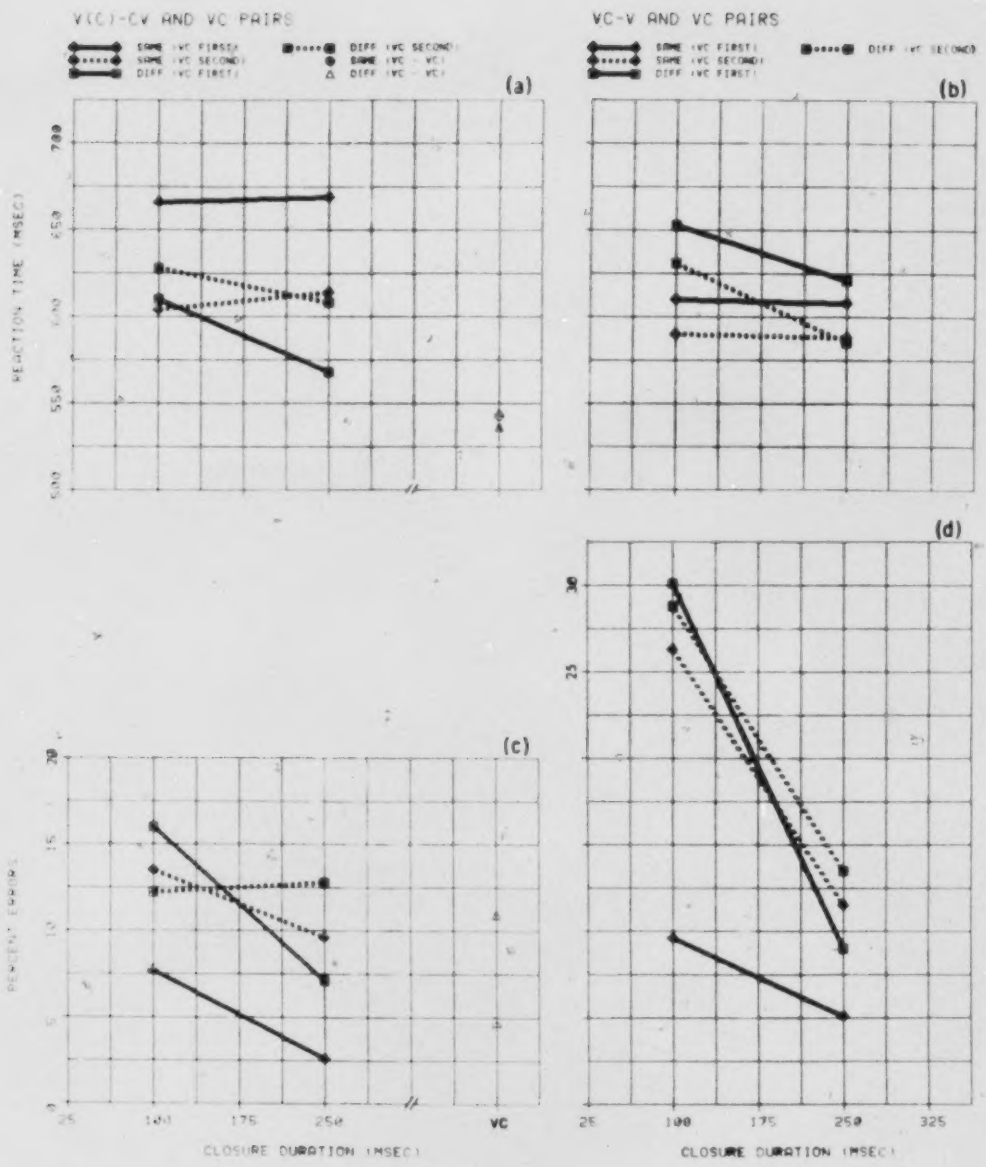


Figure 7: Mean latencies of correct responses and error rates in the same-different task: V(C)-CV and VC-V stimuli paired with VC stimuli, and VC pairs.

but since this analysis was not quite legitimate because of the highly asymmetric distribution of the error scores, detailed results will not be reported here.

The results for stimulus pairs containing VC stimuli are shown in Figure 7 with panels arranged as in Figure 6. The RTs (Figures 7a and 7b) deserve little comment, for, despite their apparent orderliness, a 5-way analysis of variance did not reveal a single significant effect. (Not even the interaction of same/different consonant with closure duration approached significance.) Latencies for VC pairs seemed to be faster than for other stimulus pairs (Figure 7a); however, this difference was not tested for significance.

It was again in the error rates that differences emerged more clearly. Figures 7c and 7d show that pairs in which the VC stimulus came first and which had identical consonants had much lower error rates than other stimulus combinations. Thus, temporal order of the stimuli in a pair clearly made a difference for "same" responses; for "different" responses, a similar effect was observed at the longer closure duration only. (Note that these effects tended to be reversed in terms of RTs; however, a true speed-accuracy tradeoff could hardly underly this inconsistency.) Except for the steep increase in errors for pairs containing VC-V stimuli at the shorter closure duration, the error patterns for the two stimulus types were quite similar. Again, no practice effects were evident. All relevant effects were significant in an analysis of variance.

Although RTs for VC pairs tended to be faster, their error rates were comparable to those for most other pairs at the longer closure duration. At the 250-msec closure duration, only pairs of VC-V stimuli had highly elevated error rates (Figure 6a): following both target consonants with irrelevant vowels introduced a strong tendency to respond "same" to consonants that actually were different, regardless of whether the two vowels were the same or not.

Discussion

As far as RTs are concerned, this experiment was not particularly successful. Inter- and intra-subject variability was too great and error rates too high to lead to useful results, apart from the marginally significant vowel effect in Task 3. However, if the view is accepted that the error rates convey very much the same information as the latencies, the relatively greater consistency of the error patterns permits one to draw conclusions that originally were to be based on the RTs. It should be noted that these conclusions apply only to relatively inexperienced listeners; so far, there is little evidence that practiced listeners are sensitive to the context following VC targets in any systematic way.[6]

The principal result is the effect of the relationship of the final vowels on "same" judgments about the stop consonants in pairs of VCV

---

[6]See Repp (1976b). I also served as a subject in the present experiment and showed no systematic effects of context (at least, no effects consistent with those shown by inexperienced listeners).

utterances. As in the earlier studies by Repp (1976a, 1976b), the effect persisted even at a relatively long closure period. In addition, the present experiment shows that it is equally present in VC-V stimuli which do not contain any explosive transitions. This rules out the hypothesis that the explosive transitions mediated the effect of the final vowel.

It will be recalled that Pisoni and Tash (1974) and Wood and Day (1975) demonstrated effects of the final vowel on judgments about syllable-initial stop consonants (explosive transitions). I hypothesized (Repp, 1975) that this effect was due to the acoustic variation of explosive transitions with the following vowel, and I demonstrated that the effect is also obtained in V(C)-CV utterances, where part of the consonantal information (the implosive transitions) is independent of the final vowel. However, the explanation was always possible that the listeners simply ignored the implosive transitions, or alternated between basing their decisions on implosive or explosive transitions, or perceptually integrated these cues because they signalled the same place of articulation. These interpretations no longer seem tenable. In VC-V stimuli, a final vowel containing no consonantal cues whatsoever biases judgments about events that are acoustically independent of it and occur as much as 250 msec earlier. This effect is of the same magnitude as that obtained in V(C)-CV stimuli, which suggests that it is of a more general nature and does not depend on the "connectedness" of portions of the signal by phonetically compatible cues. Rather, the perception of certain acoustic cues seems to be sensitive to any speech information that follows within a considerable time span. Of course, this time span will depend on a number of factors, such as the salience of the critical cue and the experience of the listener.

It is instructive to evaluate systematically the effect of adding final vowels (or CV portions) to one or both VC stimuli in a pair. The error rates for VC pairs constitute a baseline (Figure 7a). If a final vowel or CV portion is added to the second VC, a VC-VCV pair is obtained. It can be seen in Figures 7c and 7d that this addition of a vowel had little effect on error rates at the long closure period; but at the short closure period, errors increased considerably, especially on VC-V "different" trials. In comparing VCV-VC pairs (Figures 7a and 7b) with VCV-VCV pairs (Figures 6a and 6b), exactly the same manipulation -- adding a final vowel to the second stimulus-- is involved. Here the effects were more complex, because the relationship of the added final vowel to the final vowel of the first stimulus in the pair played a role. Errors at both closure durations were drastically reduced by making the second stimulus identical to the first when a vowel (or CV) portion was added that was already contained in the first stimulus. Adding a different final vowel to the second stimulus of a pair in which the consonants were the same had little effect at the longer closure duration, but reduced error rates at the shorter closure duration. This effect is interesting, for although the difference between the two stimuli increased, (especially in the case of VC-V stimuli), fewer errors were committed on "same" trials. The fact that the overall structure of the two stimuli became more similar may have been more important than the precise relationship of the final vowels, although the latter, of course, had an additional effect. When the target consonants were different, adding a final CV portion to the second stimulus increased errors slightly, while adding a final vowel only (VC-V pairs) increased errors drastically. The difference in the magnitudes

of these effects must have been due to the presence vs. absence of explosive transitions which conveyed relevant consonantal information. It was most surprising that the addition of a final vowel had such a large effect in VC-V stimuli at the 250-msec closure duration. Apart from this effect, the results may be cautiously interpreted to show two factors at work: similarity of overall stimulus structure, which played a role only at the short closure duration (suggesting that, at the longer closure duration, the two portions of each stimulus no longer formed one perceptual unit or chunk), and complete identity, which was effective at both closure durations.

Consider now the effect of adding a final vowel or CV portion to the first stimulus in a pair. Doing this to a VC-VC pair results in a VCV-VC pair. The effect is an increase in errors on "same" trials but not on "different" trials, except at the short closure duration for VC-V stimuli (Figure 7d). In each case, the manipulation eliminates the advantage of "same" trials, which apparently requires that two identical stimulus portions follow directly upon each other. (Note that the advantage of "same" trials was found in VC-VCV pairs, where no auditory information intervened between the two identical VC portions.) When a vowel or CV portion is added to the first stimulus in a VC-VCV pair, a VCV-VCV pair results in which the overall stimulus structure of the two stimuli is equal. On "same" trials, the error rates for VC-VCV pairs are more like those of VCV-VCV pairs with different vowels at the short closure duration, but like those of VCV-VCV pairs with identical vowels at the long closure duration. This again suggests that the final vowel or CV portion formed a perceptual unit with the VC portion at the shorter closure duration only. In each case, there is an advantage for two identical perceptual units following directly upon each other, be they VCs or VCVs. The effect of adding a vowel to the first stimulus on "different" trials is similar to the effect of adding a vowel to the second stimulus: a moderate increase in errors for V(C)-CV stimuli, and a large increase for VC-V stimuli, regardless of closure duration. The increase in errors at the short closure duration may also reflect a bias towards "same" responses arising from similarity in overall structure.

The data suggest, then, that the average unpracticed subject processes the stimuli as follows. All the information that occurs prior to the onset of the second VC stimulus is phonetically interpreted and stored. The information beginning with the second VC is first compared to the stored information in a holistic manner. In this holistic comparison, the size of the units compared is determined by the total information held in storage, that is, if the first stimulus was a VCV (even with a closure period of 250 msec), the units to be compared will be VCVs, if it was a VC, they will be VCs. (In the latter case, if the second VC is followed by further information after a relatively short interval, the listener may have difficulty in segregating the VC portion as a unit for comparison.) If the second unit exactly matches the first unit held in storage, an accurate (and fast) "same" response is issued. The low error rates for identical VC-V stimuli with a short closure period suggest that these matches take place at a prephonetic (auditory) level; otherwise, there should have been more errors on "same" trials because of the high uncertainty about the phonetic identity of these stimuli (cf. Experiment I). If the holistic match is negative (or already while it is being performed), a more analytic comparison is conducted, most likely between phonemic stimulus representations. The final vowel

in VCV stimuli can be ignored in this comparison, but it may still have indirect effects on the identification of the stop consonant. These indirect effects may account for part of the error pattern, such as the striking difference between VC-V/VC-V "same" (vowel different) trials and VC-V/VC "same" trials at the short closure duration.

Clearly, this is not a full account of what is going on in the listener's head in this complex task. Various individual differences and strategies may be involved. But the availability of a special prephonetic mode of comparison for the detection of exact identity seems fairly clear. The good agreement with the results of Repp (1976b) should be noted, in terms of error rates, at least. There was no consistent effect of closure duration on RTs when the second stimulus was of the V(C)-CV type. This also agrees with the earlier results and indicates that the subjects did not make their decisions solely on the basis of the explosive transitions. Clearly, however, the explosive transitions were taken into account, as shown by the difference in "different" error rates between V(C)-CV and VC-V pairs. Further research yielding cleaner RT data will be needed to gain more insight into the precise processing strategies employed by listeners in this task.

EXPERIMENT III

This experiment investigated the perceptual interaction between implosive and explosive transitions in VCV stimuli by a new method: systematic manipulation of the acoustic structure of the transitions. Consider a VCV utterance with a short closure duration (for example, 25 msec). The medial stop is almost always perceived according to the explosive transitions, even if the implosive transitions are appropriate for a different place of articulation (cf. Dorman et al., 1975, and the present Experiment I). In other words, both /ab-dɛ/ and /ad-dɛ/ are perceived as /adɛ/, and both /ab-bɛ/ and /ad-bɛ/ are perceived as /abɛ/, if the closure period is made sufficiently short. What happens if the explosive transitions are chosen so that the second syllable is ambiguous between /bɛ/ and /dɛ/ when presented in isolation? Will it be equally ambiguous when preceded by /ab/ or /ad/ at a short closure duration? Or will the (unambiguous) implosive transitions determine the phonetic percept in this case? Their effect could be either assimilative or contrastive; because of the close contiguity of the interacting transitions, and since the implosive transitions are not perceived as a separate phonemic event, an assimilative effect seems more likely. Such an effect would provide evidence of perceptual integration of implosive and explosive transitions, while absence of any effect would support a perceptual interruption hypothesis (Massaro, 1975) or at least suggest that implosive transitions play no perceptual role at very short closure durations.

Consider now the reverse case. As the closure duration is increased, a stimulus like /ab-dɛ/ changes perceptually from /adɛ/ to /ab-dɛ/. At comparable closure durations, /ad-dɛ/ remains /adɛ/ in perception; geminate consonants (/ad-dɛ/) are perceived only at much longer closure durations (Repp, 1976b). What happens if the implosive transitions are made ambiguous between /ab/ and /ad/? When followed by /dɛ/ at an intermediate closure duration (115 msec, say), will the perceptual result be /adɛ/ or /ab-dɛ/? When followed by /bɛ/, will it be /abɛ/ or /ad-bɛ/? Again, the effect of the explosive transitions on the perception of the ambiguous implosive transi-

tions could be either assimilative or contrastive, or absent altogether. A prediction is much more difficult to make in this case. Again, an assimilative effect would provide evidence of perceptual integration over a period as long as the closure duration used.

The method used was to construct acoustic continua of implosive transitions (/ab/-/ad/) and explosive transitions (/bɛ/-/dɛ/) and to investigate shifts in the phoneme boundaries on these continua as a function of the phonetic identity of the preceding (following) transitions. Four control conditions were included. In two of them, the VC and CV portions were presented in isolation. In the other two, the VC-CV combinations had a closure duration of 265 msec, so that the implosive transitions were always perceived as a separate phonemic event, even when phonetically compatible with the explosive transitions. (The single-geminate boundary lies around 213 msec -- Repp, 1976b). If there is any perceptual interaction between implosive and explosive transitions over this long temporal distance, it is most likely contrastive. A rating scale was used to judge the stimuli, since it was thought possible that the perceived clarity of a consonant might be affected by preceding (or following) compatible (or incompatible) transitions, independently of its perceived identity.

Method

Subjects. Ten new volunteer subjects participated. I also served as a subject, but my data were not combined with those of the other subjects.

Stimuli. All stimuli were prepared on the Haskins Laboratories parallel formant synthesizer. Two stimulus continua were constructed: a VC continuum of seven syllables ranging perceptually from /ab/ to /ad/, and a CV continuum of seven syllables ranging perceptually from /bɛ/ to /dɛ/. The stimuli within each continuum differed only in the offset (onset) frequencies and trajectories of the second- and third-formant transitions, spaced in equal steps between the two endpoint stimuli. The stimuli were selected so that the phoneme boundary would fall approximately in the center of each continuum. The VC stimuli were 185 msec long, with 35-msec transitions; the CV stimuli were 300 msec long, with 50-msec transitions (as in the previous experiments). The VC stimuli all had the same F_1 transition as the /ab/ stimuli in Experiment I and earlier experiments (unlike the stimuli in Experiment II).

Two stimulus tapes were prepared. The CV tape first contained a random series of 75 CV syllables consisting of the seven CV stimuli with the following frequency distribution: 5 times (1,2,3,3,3,2,1). This distribution of stimuli was used to provide more reliable information in the region of the phoneme boundary and was maintained in all other conditions. The initial CV series was followed by a series of 150 stimuli consisting of the same CVs preceded by either /ab/ or /ad/, the two endpoint stimuli of the VC continuum. The closure interval was 25 msec. Another analogous series of 150 stimuli followed, with a closure period of 265 msec. These sequences were arranged in successive blocks of 30 stimuli, each containing one cycle of all stimulus combinations, with the basic stimulus frequency distribution described above.

The VC tape was exactly analogous. An initial 75-item VC series was followed by two 150-item VC-CV series in which each VC stimulus was followed by either /bɛ/ or /dɛ/, the two endpoint stimuli of the CV continuum. The closure period was 115 msec in the first series and 265 msec in the second. The CV and VC tapes had identical stimulus randomizations, with reversed roles of the VC and CV portions.

Procedure. All subjects received the conditions in the same order: first the CV tape, then the VC tape, and the stimulus sequences in the order described above. In the initial CV series, the subjects were instructed to rate each consonant on a scale ranging from 1 to 6, where 1 represented a "very clear B", 3 "ambiguous, more like a B", 4 "ambiguous, more like a D", and 6 a "very clear D". Subjects were urged to use the extreme ratings at least occasionally, that is, to make their judgments according to the relative goodness of the stimuli and not according to how they compared with real speech. The subjects were exposed to a portion of the stimulus series before actually beginning the task. In the following conditions, the subjects were asked to maintain the criteria established during the initial series, that is, to give generally poorer ratings if all stimuli sounded poorer and generally better ratings if all stimuli sounded better. For the 25-msec CV condition, the subjects were merely told that each CV syllable would be preceded by the vowel /a/; nothing was mentioned about the implosive transitions. For the 265-msec CV condition, the subjects were told that each CV syllable would be preceded by either /ab/ or /ad/. These initial syllables were to be ignored, and only the relative category goodness of the initial consonant of the second syllable was to be evaluated.

In the VC conditions, the subjects first rated the syllable-final consonants on the same six-point scale. Then, in the 115-msec condition, a different response mode was introduced because of the perceptual heterogeneity of the stimuli (either one or two intervocalic consonants). Instead of using the rating scale, the subjects wrote down "1" when they heard a single consonant (/abɛ/ or /adɛ/) and "2" when they heard two different consonants (/ab-dɛ/ or /ad-bɛ/). Finally, in the 265-msec condition, the rating scale was used again to evaluate the first (syllable-final) consonant, ignoring the /bɛ/ or /dɛ/ that followed.

The equipment was the same as in previous experiments. All conditions were administered in a single session of about one hour.

Results

The results of the CV conditions are shown in Figures 8a and 8b. The dashed lines represent the ratings for CV stimuli in isolation. The other two functions in each panel of Figure 8 represent responses to CV syllables preceded by /ab/ and /ad/, respectively.

It is obvious that the VC precursors had an effect in the 25-msec condition but not in the 265-msec condition. The former effect was assimilative, as expected, and remarkably consistent from subject to subject, as reflected in its high significance ($F_{1,9} = 45.87$, $p << .01$). The significance test was performed on the difference between the effects of the two precursors on the ratings; the control data (CVs in isolation) were not

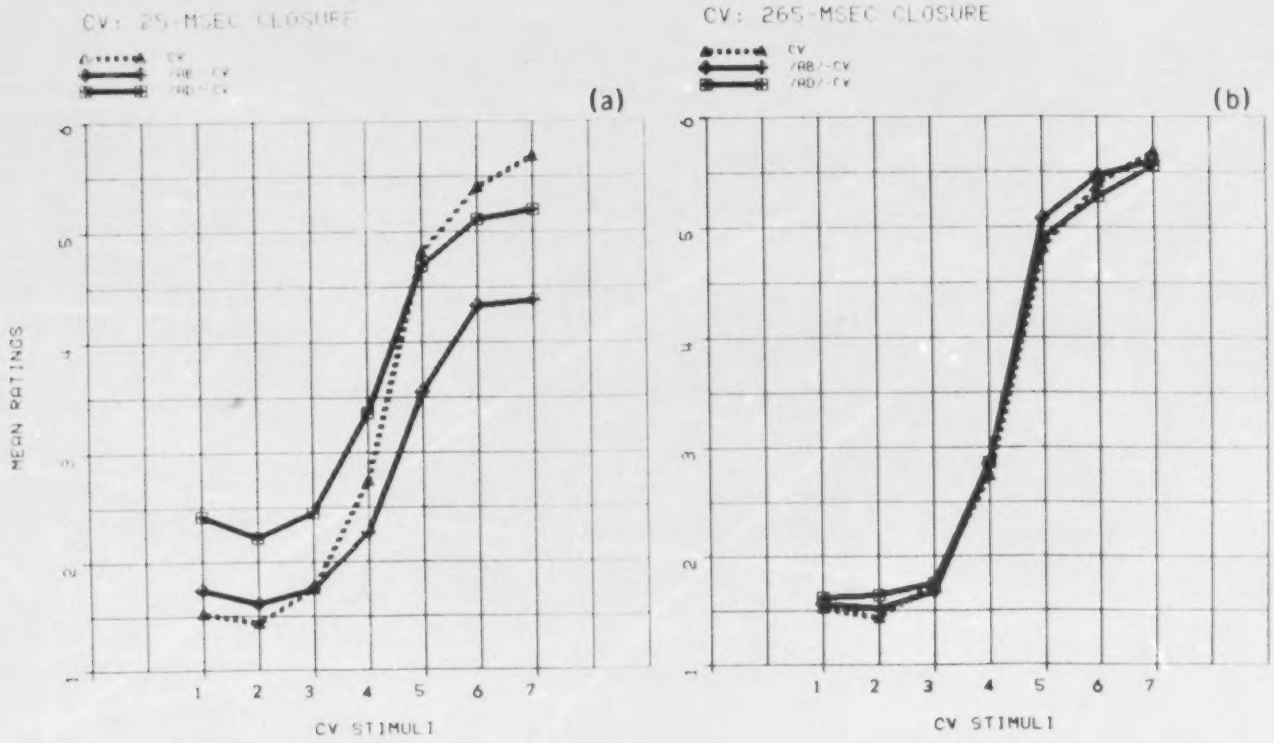


Figure 8: Mean "D-ness" ratings of CV stimuli from the /bɛ/-/dɛ/ continuum in isolation and when preceded by either of two VC precursors, at two closure intervals. (The dashed functions in the two panels are based on the same data.)

FIGURE 8

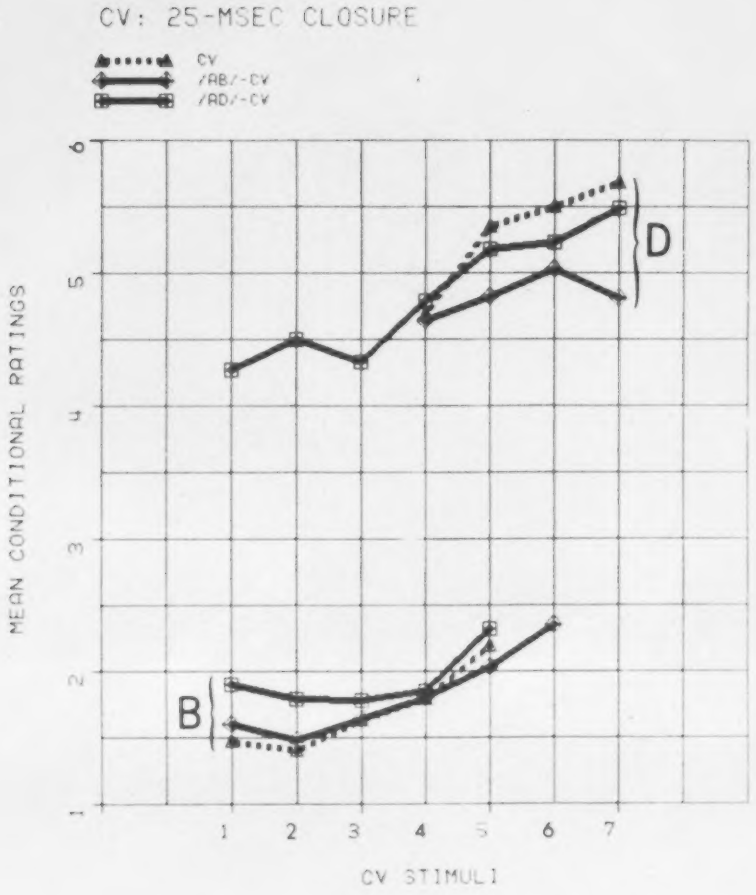


Figure 9: The data of Figure 8a, plotted separately for ratings falling between 1 and 3 (B) and between 4 and 6 (D).

included in this analysis. The precursor effect did not interact with position on the stimulus continuum; it can be seen in Figure 8a that it was equally present for each of the seven CV stimuli. The rating functions for the VC-CV stimuli were not only flatter than the function for CV stimuli in isolation, but they also reached an earlier asymptote at one end. When the same results were plotted in terms of the percentage of "D" responses (that is, the percentage of ratings falling between 4 and 6), the pattern was identical. This suggests that the implosive transitions simply "got through" on a percentage of trials. This was not entirely unexpected; at a closure duration of 25 msec, the perceptual dominance of the explosive transitions is not perfect (Dorman et al., 1975, and the present Experiment I). One subject actually heard /adɛ/ whenever the VC portion was /ad/. Another subject reported hearing /ab-dɛ/ on a number of trials.

Thus, the question arises whether the effect of the VC precursor was all-or-none or gradual in nature. Did it consistently bias the perception of the CV portion, or did it just intrude on a certain small number of trials and have no effect on all others? One way of answering this question is to make the average ratings conditional on whether they fell between 1 and 3 (B) or between 4 and 6 (D). These conditional ratings for the 25-msec condition are shown in Figure 9. Only data points with at least 10 responses in the relevant category are shown. The entries represent means calculated over all individual responses of all subjects, that is, different subjects contributed different numbers of responses, and therefore no statistical analysis could be conducted. It is evident from Figure 9 that the precursor effect was reduced in terms of conditional ratings, but a smaller effect in the predicted direction clearly remained. In other words, /bɛ/ preceded by /ad/ was indeed perceived as a "poorer B" than /bɛ/ preceded by /ab/ or by silence, and /dɛ/ preceded by /ab/ was perceived as a "poorer D" than /dɛ/ preceded by /ad/ or by silence. We may conclude, then, that the VC precursor exerted a genuine biasing effect on the perception of the explosive transitions on most or all trials. Note, however, that preceding an unambiguous CV with a phonetically compatible VC precursor did not improve its ratings compared to the same CV syllable in isolation; thus, there was no positive contribution of the implosive transitions to the perceived clarity of the consonant.

It is curious that I was the only listener who showed a precursor effect in the opposite direction, that is, a contrast effect, although I never perceived more than a single consonant in the 25-msec condition. It is not clear why my extensive experience with the stimuli should have led to this surprising reversal.

The obvious absence of any average precursor effect in the 265-msec CV condition (Figure 8b) may not be representative of individual listeners. Of the ten subjects, two showed assimilation effects, five showed contrast effects, and the remaining three showed irregular effects or none at all. I showed an assimilation effect. Thus, although some of these effects may just represent random variation, it seems that the VC precursor did affect the perception of the CV syllables, but, in different directions for different listeners. At present, the basis of the individual differences is obscure.

The results of the VC rating conditions are shown in Figure 10. Figure 10a shows the 115-msec condition. Here, the responses indicated the number of consonants heard. The ordinate is labeled percent "D" responses, which means the percentage of ratings between 4 and 6 for VC syllables in isolation, the percentage of "2" responses for stimuli followed by /bɛ/, and the percentage of "1" responses for stimuli followed by /dɛ/. Plotted in this way, it is evident that the two CV "postcursors" had little differential effect. Again, however, individual results varied widely -- more so than one would expect from mere random variability. Four subjects were more likely to hear one consonant with the /bɛ/ postcursor than they were to hear two consonants with the /dɛ/ postcursor, one subject showed the opposite effect, and the remaining subjects showed different effects in different regions of the VC continuum. Such an interaction is weakly evident also in Figure 10a: at the /ab/-end of the VC continuum, the /bɛ/-function lies above the /dɛ/-function, and this relationship is reversed as the /ad/-end of the VC continuum is approached. Seven out of ten subjects showed results at least partially compatible with this pattern, which, however, is not readily interpretable and was not statistically significant.

Much more consistent than the differences between the two postcursors was the difference between the postcursor functions and the function for VC syllables in isolation ($F_{1,9} = 5.9$, $p < .05$, for the main effect; $F_{4,36} = 13.3$, $p < .01$, for the interaction with position on the continuum).[7] The difference can be broken down into two components: lower asymptotes of the postcursor functions (at least at the /ab/ end of the VC continuum), and a general shift in the VC category boundary towards the /ad/ end when a postcursor followed. No matter which CV syllable followed, the VC portion was more likely to be perceived as /ab/ than in isolation. The reason for the first component was probably general uncertainty due to the relative difficulty of the task. The reason for the second component is not clear, except that it is reminiscent of the general difficulties subjects had in perceiving /ad/ correctly in earlier experiments.

I again produced a curious result in the 115-msec condition: I needed a while to hear any instances of two consonants at all, which made my data quite useless. (The same happened in a replication of the experiment.) Warm-up effects of this sort may have played a role with some of the other subjects, too, although they seemed to have much less trouble.

Finally, the results of the 265-msec VC condition need to be discussed. They are shown in Figure 10b. (The data of one subject had to be excluded in this condition because he apparently responded to the CV portions of the stimuli.) It can be seen that there was a small postcursor effect in the predicted direction, that is, a contrast effect. Slight contrast effects were shown by five subjects and myself; the remaining subjects showed no systematic effects. No listener showed any assimilation effect in this condition. Due to this relative consistency between subjects, the postcursor

---

[7] In the statistical analysis, the seven positions were reduced to five by combining the two positions at each end of the continuum, so that an equal number of observations was available at each of the resulting five positions.

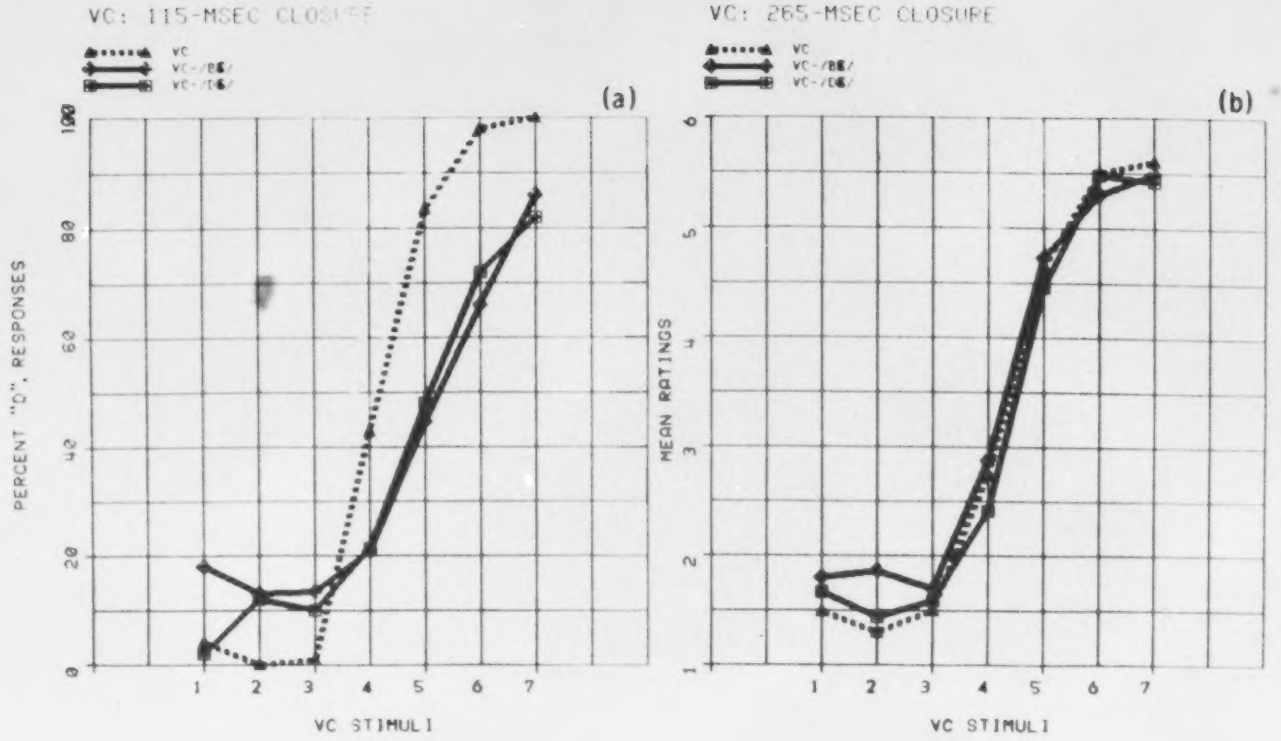


FIGURE 10

Figure 10: Percentage of "D" responses (panel a) and mean "D-ness" ratings (panel b) for VC stimuli from the /ab/-/ad/ continuum in isolation and when followed by either of two CV postcursors, at two closure intervals. (The dashed functions in the two panels are based on the same data but plotted differently.) Percentages of "D" responses in panel a are inferred from "1" and "2" responses and ratings given in the task (see text).

effect reached significance ($F_{1,8} = 7.97$, $p < .05$), although most individual effects were smaller than those in the 265-msec CV condition. The effect did not interact significantly with position on the VC continuum. Again, there was no evidence of any increase in the perceptual clarity of an unambiguous VC syllable when followed by a CV syllable.

Discussion

These results demonstrate that implosive and explosive transitions are not perceived independently of each other. At short closure durations, the implosive transitions are perceptually dominated by the explosive transitions and only a single consonant is heard. Nevertheless, the implosive transitions bias the perception of the explosive transitions. The conditional ratings (Figure 9) indicate that this is, at least in part, a genuine perceptual bias due to perceptual integration of auditory or phonetic information. Part of the effect may also be due to occasional perceptual dominance of implosive over explosive transitions. Whether the perceptual integration occurs at an auditory or at a phonetic level is not clear at present. This issue could be further investigated by varying the acoustic structure of the VC precursor within phonetic categories.

When the closure period is lengthened, the implosive transitions emerge as a separate phonemic percept if they are incompatible with the explosive transitions. As the results of the 115-msec condition show, the nature of this percept is not consistently influenced by the identity of the postcursor. However, the mere presence of a CV postcursor biased the perception of the VC portion towards labials. This effect can no longer be due to straightforward perceptual integration, but it probably represents some more general perceptual interaction as exemplified also in Experiment I, Task 1 (VC-V stimuli). In terms of Massaro's (1975) theory, the results may be interpreted to indicate that /ad/ required more processing time than /ab/, so that a following event interfered more with the former than with the latter.

When separated by a closure period of 265 msec, the perception of implosive and explosive transitions is largely independent, but there is a tendency towards small contrastive effects that, surprisingly, are more consistent in the backward direction than in the forward direction. This may reflect the lower perceptual salience of implosive transitions. Although the present VC stimuli were as consistently identified as the CV stimuli in isolation, their susceptibility to contextual factors seemed to be greater, perhaps due to the absence of a "protective" continuation of the signal (such as a release burst might provide it). The contrastive postcursor effects are evidence that, at least occasionally, phonetic decisions about the implosive transitions are postponed until events occurring as much as 265 msec later have been phonetically interpreted. Of course, it may be the normal mode of processing speech to phonetically recode chunks of VCV size or larger. This agrees well with the results of Experiment II and the earlier RT studies.

REFERENCES

Dorman, M. F., Raphael, L. J., Liberman, A. M., and Repp, B. H. (1975) Maskinglike phenomena in speech perception. Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.

Liberman, A. M. (1975) How abstract must a motor theory of speech perception be? Haskins Laboratories Status Report on Speech Research SR-44, 1-15.

Malmberg, B. (1955) The phonetic basis for syllable division. Studia Linguistica 9, 80-87.

Massaro, D. W. (1975) Preperceptual images, processing time, and perceptual units in speech perception. In Understanding Language, An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics, ed. by D. W. Massaro. (New York: Academic Press), 125-150.

Pisoni, D. B., and Tash, J. (1974) "Same-different" reaction times to consonants, vowels, and syllables. In Research on Speech Perception, Progress Report No. 1 (Dept. of Psychology, Indiana University), 129-139.

Repp, B. H. (1975) "Coperception": A preliminary study. Haskins Laboratories Status Report on Speech Research, SR-42/43, 147-157.

Repp, B. H. (1976a) Coperception: Two further preliminary studies. Haskins Laboratories Status Report on Speech Research, SR-45/46, 141-152.

Repp, B. H. (1976b) Perception of implosive transitions in VCV utterances. Haskins Laboratories Status Report on Speech Research, SR-48, 209-233.

Wood, C. C., and Day, R. S. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percep. Psychophys., 17, 346-350.

Phonetic Recoding and Reading Difficulty in Beginning Readers

Leonard S. Mark,† Donald Shankweiler,† Isabelle Y. Liberman,† and Carol A. Fowler††

ABSTRACT

The results of a recent study (Liberman, I., Shankweiler, Liberman, Fowler, and Fischer, 1977) suggest that good beginning readers are more affected than poor readers by the phonetic characteristics of visually presented items in a recall task. The good readers made significantly more recall errors on strings of letters with rhyming letter names than on nonrhyming sequences; in contrast, the poor readers made roughly equal numbers of errors on the rhyming and nonrhyming letter strings. The purpose of the present study was to determine whether the interaction between reading ability and phonetic similarity may be solely determined by different rehearsal strategies of the two groups. Accordingly, good and poor readers were tested on rhyming and nonrhyming words using a recognition memory paradigm that minimized the opportunity for rehearsal. Performance of the good readers was more affected by phonetic similarity than was that of the poor readers, in agreement with the earlier study. The present findings support the hypothesis that good and poor readers do differ in their ability to access a phonetic representation.

INTRODUCTION

Many investigators see the root cause of reading disability in school children as a deficit in perceptual learning (for example, Bender, 1957; Frostig, 1963; Silver and Hagin, 1960). Their research has emphasized the importance of visual processes such as those involved in the identification of letter shapes and the scanning of text. However, critical surveys of such

---

†Also University of Connecticut, Storrs.

††Also Dartmouth College, Hanover, New Hampshire.

Acknowledgment: This paper is based on a Masters thesis by the senior author. The investigation was supported by grants to Haskins Laboratories from NICHD (HD-01994) and a training grant to the University of Connecticut from NICHD (HD00321-04). The authors wish to express their appreciation to Leonard Katz, Sandra Prindle, Michael Turvey, and Robert Verbrugge for their advice and criticisms on earlier drafts of this manuscript and to Michele Werfelman for her assistance in the collection of data.

[HASKINS LABORATORIES Status Report on Speech Research SR-49 (1977)]

research (Benton, 1962, 1975; Hammill, 1972; Vernon, 1960) produced little hard evidence to support the hypothesis that visual and directional factors figure heavily in most cases of reading disability. This conclusion was reaffirmed by the work of Shankweiler and Liberman (1972), Vellutino, Steger, and Kandel (1972), Vellutino, Pruzek, Steger, and Meshoulam (1973), and Vellutino, Steger, Harding, and Phillips (1975).

In view of the repeated failure to establish visual-perceptual deficits as a major problem in learning to read, several investigators have begun to examine other cognitive prerequisites for reading acquisition, in particular, those relating to the child's primary language abilities. These investigations (for example, Bloomfield, 1942; I. Liberman, 1971, 1973; Mattingly, 1972; Rozin and Gleitman, 1977; Shankweiler and Liberman, 1976) have suggested that reading should not be viewed as an independent ability, but as parasitic upon the spoken language. If reading is a derivative of speech and acquired by the child only after he has acquired speech, it is reasonable to consider how learning to read may build upon the earlier language acquisitions of the young child.

Although both good and poor readers speak and understand the language, it may be that poor readers have deficiencies in certain subtle aspects of language development that are not evident even to trained observers. The present research examines this possibility. Specifically, its purpose is to explore the role of phonetic recoding in reading acquisition and to investigate the hypothesis that good and poor beginning readers differ in their ability to access and to use a phonetic representation.

A notable characteristic of language is that the meaning of the longer segments (for example, sentences) transcends the meaning of the shorter segments (for example, words); it follows that a listener would have to maintain the smaller units in some temporary store, until a sufficient number of them have accrued--to enable him to apprehend the meaning. It has been argued (A. Liberman, Mattingly, and Turvey, 1972) that a phonetic representation is used for this purpose and that it is uniquely suited to the short-term storage requirements of language. Our own research has emphasized two additional functions of the phonetic representation of spoken language (Shankweiler and Liberman, 1976; I. Liberman, Shankweiler, Liberman, Fowler, and Fischer, 1977). We have speculated that a language user may employ a phonetic representation in order to access his mental lexicon and to reconstruct the prosodic information that is crucial to understanding speech. We have also suggested that readers of a language may continue to use a phonetic representation, just as hearers do, rather than develop a new mode of processing for the written language.

There is considerable experimental evidence to support the view that people do employ a phonetic code to store visually presented letters or words, even under circumstances where it is disadvantageous to do so (for example, Conrad, 1964, 1972; Baddeley, 1966, 1968, 1970; Hintzman, 1967; Kintsch and Buschke, 1969). Typical studies presented subjects with letter or word sequences to be read silently and then recalled. The investigators usually reported that most confusion errors were based on the sound of the letter or word, rather than on its visual appearance.

In addition to these considerations, there is reason to believe that phonetic recoding is of special significance for the beginning reader who is learning how the alphabet works. Consider the relationship between the alphabet and the spoken language. English, unlike the logographic writing system of Chinese and the Japanese Kanji, uses a symbol system, the alphabet, that is keyed largely to the sound structure of the language. If the child has learned something about how the spelling reflects the sound structure, he will be able to offer at least an approximate pronunciation of new words. However, to take full advantage of the benefits inherent in the symbol economy of an alphabet, the reader must be able to employ an analytic strategy, grouping the letter segments into articulatory units and mapping them into speech, rather than treating words as irreducible wholes (Shankweiler and Liberman, 1976; Liberman et al., 1977).

However, in order to use an analytic strategy, the reader must recognize that the alphabet is largely a direct representation of the phonemes in speech. Whereas the recognition of two spoken utterances like bet and best as different words, is sufficient for the comprehension of these as lexical items, the process of mapping the written word onto its spoken counterpart requires, in addition, recognition of the number and identity of the phonemes contained in the spoken word. There is now considerable evidence to suggest that the ability to recognize phoneme segments in speech is a predictor of success in learning to read (Savin, 1972; Helfgott, 1976; Liberman et al., 1977; Zifcak[1]).

In view of the evidence that poor readers have difficulty in performing phoneme segmentation tasks, it is appropriate to ask whether poor readers are also deficient in the ability to construct and employ a phonetic representation. Conceivably, poor readers might attempt to retain script as shapes, rather than as phonetic entities. Using a recall-memory task, our research group has found evidence to suggest that good and poor readers do differ in their phonetic coding ability (Liberman et al., 1977). In that study, good and poor second grade readers were presented with sequences of letters for recall. Half of the sequences were composed of rhyming consonants (from the set B C D G P T V Z), the remainder of nonrhyming consonants (from the set H K L Q R S W Y). Each of the strings of five upper-case letters was displayed tachistoscopically for three seconds. The subjects were instructed to print as many of the letters as they could remember, either immediately after presentation or after a 15-sec delay. Their responses were scored both with and without regard to serial position.

Under both recall conditions, the good readers displayed significantly more phonetic interference than the poor readers, as measured by the differences in total errors between the rhyming and nonrhyming sequences. Because of this interaction between reading ability and phonetic similarity, the difference in performance between good and poor readers cannot be explained by supposing that the two reading groups differ in "general memory capacity." The differences also cannot be attributed to a serial-ordering

---

[1]M. Zifcak, Phonological awareness and reading acquisition in first grade children. Unpublished doctoral dissertation, University of Connecticut (in preparation).

problem in the poor readers, since the effects were significant even when recall was scored without regard to serial position.

It appeared, then, that the phonetic characteristics of the letter names had a differential effect on recall in good and poor readers. From this, it was assumed that the good readers are better able to access and use a phonetic representation in short-term memory than the poor readers. An alternative interpretation, however, would ascribe these findings to differences in rehearsal strategy for the two reading groups.[2] If the poor readers were able to rehearse fewer letters than the good readers before recall began, the rhyming letters would have less opportunity to interfere. This might give rise to the pattern of results obtained: inferior recall of the nonrhyming items by the poor readers, but little difference between the groups on the rhyming letters.

The present experiment was undertaken primarily in an effort to resolve this ambiguity. A paradigm originally devised by Hyde and Jenkins (1969) for a different purpose was adapted for this study, because it permits us to test memory in a way that minimizes the opportunity for rehearsal. The procedure involves a test list of words followed by a recognition list. The subjects are not informed at the time of the presentation of the first list that a subsequent test of recognition memory will follow. Thus, the task appeared to the child merely as a reading task. If differential rehearsal rates were responsible for the earlier results, then differences in phonetic similarity should disappear with this new procedure. However, should the findings of the present study replicate those obtained in the previous research, there would be support for the interpretation that the poor readers have a deficit in accessing or using a phonetic representation derived from script.

A second reason for undertaking the present study was to test the phonetic coding ability of the two groups of readers in a task more nearly resembling a realistic reading situation. This was accomplished by using words, rather than letter strings, as the stimulus items.

METHOD

Subjects

The subjects were second grade school children in the Mansfield, Connecticut public school system. Children were selected for pretesting on the basis of their total reading grade on the Stanford Achievement Test (SAT), that had been administered by the schools during the fourth month of the school year. In this preliminary screening, children with total reading grades between 3.5 and 5.0 on the SAT were candidates for the good reading group, while those with reading scores between 1.5 and 2.4 were considered for the poor reading group. Final selection of the two reading groups from among these children was made in the seventh month of the school year by administering the word recognition subtest of the Wide Range Achievement Test (WRAT) (Jastak, Bijou, and Jastak, 1965). The criterion for inclusion in the good reading group was a WRAT grade level between 3.1 and 5.0. A child was

[2] R. Crowder: personal communication.

selected for the poor reading group if his WRAT grade level was in the range of 1.5 to 2.4.

Thirty-seven children (19 good readers and 18 poor readers) met the WRAT criteria for participation in the experiment. Seven subjects (four good and three poor readers) had to be dropped because their data were incomplete due to an experimenter error. Another poor reader had to be excused from the experiment because he was unable to read more than 50 percent of the words on the recognition list (see Scoring Method). Thus, the data analysis was based on the performance of 15 good readers with a mean WRAT grade level of 3.97 (range: 3.1 to 4.5) and 14 poor readers with a mean WRAT grade level of 2.19 (range: 1.5 to 2.4).

The good readers had a mean age of 92.4 months, while the mean age of the poor readers was 94.0 months [$t(27) = .97$, $p < .40$]. The relative intelligence (IQ) of the two reading groups was assessed by the Wechsler Intelligence Scale for Children, Revised Edition (Wechsler, 1974). The good readers had a mean Full Scale IQ of 114.2 (Verbal Scale IQ = 113.1, Performance Scale IQ = 112.5). The Full Scale, Verbal, and Performance IQ means for the poor readers were 109.0, 106.4, and 110.9 respectively. The intelligence scores of the two reading groups did not differ significantly on any of the three scales: Full Scale, $t(27) = 1.05$, $p < .40$; Verbal, $t(27) = 1.52$, $p < .20$; Performance, $t(27) = .29$, $p < .80$.

Word Lists

The word lists consisted of monosyllables chosen from Part One of the Cheek Master Word List (Cheek, 1974). The words (see Table 1) were limited to the first grade level (1.0 - 2.0) in order to ensure that the poor readers could read the bulk of the words presented, despite their reading handicap.

The initial list was composed of 28 words. The recognition list included the 28 words on the initial list and an equal number of words, the foils, not present on that list. Fourteen of the foils were phonetically paired with a word on the initial list. These are the phonetically similar (that is, rhyming) items. Word pairs were classified as phonetically similar if they met both of the following criteria: (1) they must share the same vowel sound; (2) they can differ by no more than three consonantal phonetic features in the set of "place", "manner", "voicing" and "nasality" (Wickelgren, 1966). If a set of two words failed to meet either or both requirements, they were considered to be phonetically dissimilar.

The phonetically similar foils, additionally, had to meet the requirement that they be as different as possible in visual configuration from all words on the initial list (for example, my-high, know-go). The decision to make this requirement was motivated by the possibility that some subjects might be responding primarily to the visual appearance of the word, thereby potentially confounding the results. The remaining 14 foils were both phonetically and visually dissimilar to words on the recognition list.

Given the constraint of having to select words from a first grade reading list, it was impossible to maintain strict criteria for visual dissimilarity. However, it was important to have some measure of the

TABLE 1: List of Phonetically Similar Word Pairs and Phonetically Dissimilar Words

Phonetically Similar Word Pairs

Old	Foil
know	go
my	buy
cry	high
good	could
they	way
but	what
gum	come
shoe	two
new	do
bird	word
your	for
said	red
run	done
door	more

Dissimilar Words

Old	Foil
year	best
life	guess
each	as
walk	ride
help	our
keep	did
not	cake
see	duck
friend	oh
up	off
jump	box
told	bring
yes	face
gave	brown

relative visual similarity of the two foil types to words on the initial list, so that possible visual coding strategies would not confound the results. Accordingly, several informal criteria of visual similarity were followed: (1) the two words had the same number of letters; (2) the initial letters in the words were the same; (3) the initial letters in the words were of the same shape (see below); (4) the final letters in the words were the same shape.

In the following chart, the lower-case letters are grouped into four categories reflecting "similar shape" according to a scheme devised by the authors.

Lower Case Letter Shapes

a. short curved - c o e a s m n r u
b. short straight - v w x z i
c. tall above line - h d b f l t k
d. tall below line - p q g j y

A visual-similarity matrix was constructed to compare each foil word with each word from the initial list. The numbers entered in a particular cell indicated the dimensions of visual similarity shared by a particular word-pair. The relative visual similarity of the two foil types to the words on the initial list was computed by taking the total number of times each of the four criteria was satisfied for each foil; thus, four totals were obtained for each foil word. Separate t-tests were performed on the four visual similarity measures derived for the two types of foils. No t-test was significant beyond the .05 level. This suggests that the two sets of foils were roughly comparable in visual similarity to words on the initial list.

Some words had more than one rhyming counterpart (for example, my-high, cry-buy). As a result, some foils were phonetically similar with a second word on the initial list. This somewhat undesirable situation arose with the need to increase the size of the word list, which was constrained by the limits of a first grade reading list.

Words with phonetically similar foils were equally distributed in each half of the initial list. Each half of the recognition list contained an equal number of words from four sets: phonetically-similar old words, phonetically-dissimilar old words, phonetically-similar foils, and phonetically-dissimilar foils. In addition, half of the rhyming foils preceded their rhyming counterparts from the initial list, while the remaining foils appeared after their counterparts from the initial list.

The words were hand-printed in lower case on white, three-by-five cards, using a black, felt-tipped pen. The short letters were 1/4 inch high, the tall letters 1/2 inch high.

Procedure

The children were assigned at random to one of two examiners who tested them individually.

Initial list. At the start of the experiment, the child was told that some words were going to be shown to him one at a time. He was instructed to read each word aloud and then to wait until the next word was shown. Each word was shown for as long as it took the child to pronounce it. If the child read the word incorrectly, the experimenter indicated this on the scoring sheet; no attempt was made to correct the child. However, if the child corrected himself spontaneously, the word was scored as having been read correctly.

Recognition list. After completing the initial list, the child was informed that he was going to be shown a second list of words, one at a time. (No mention of this had been made previously.) His task was to read each word aloud and then to say "yes" if he believed the word was on the old list or "no" if he believed it was not. The experimenter recorded both the child's recognition response ("yes" or "no") and whether the child read the word correctly. Before presentation of the recognition list, the examiners verified the child's comprehension of the instructions.

Scoring Method

Reading errors. Any word that was misread on either list was excluded from analysis of that child's recognition judgments. If the child misread a word on the initial list that rhymed with a foil on the recognition list, the recognition response to the phonetically similar foil was also discarded, except in cases where the foil rhymed with another word on the initial list (see previous section). These exclusions were necessary in order to ensure that errors in recognition judgments could be attributed with confidence to phonetic similarity with a word on the initial list. Any child who misread more than 50 percent of the words on the recognition list was dropped from the experiment.

Recognition judgments. A child's recognition performance on each of the four word sets was expressed as a ratio of the number of recognition errors to the total number of words read correctly in each set.

RESULTS

If the findings of Liberman et al. (1977) can be taken to reflect differences between superior and poor readers in phonetic recoding, then we may expect the following results in the present study: the good readers should make significantly more recognition errors on the rhyming foils than on the nonrhyming foils; the poor readers, on the other hand, should generate approximately equal frequencies of errors on the two types of foils. If, however, both reading groups make equal numbers of errors on each foil type, then we may suppose that opportunity for rehearsal, which was a feature of the previous investigation but not of the present one, may have accounted for the interaction between reading ability and phonetic similarity reported earlier.

Recognition Judgments

Two types of recognition errors will be considered. Of primary interest are the "false positive" errors: the child reports a word as having occurred

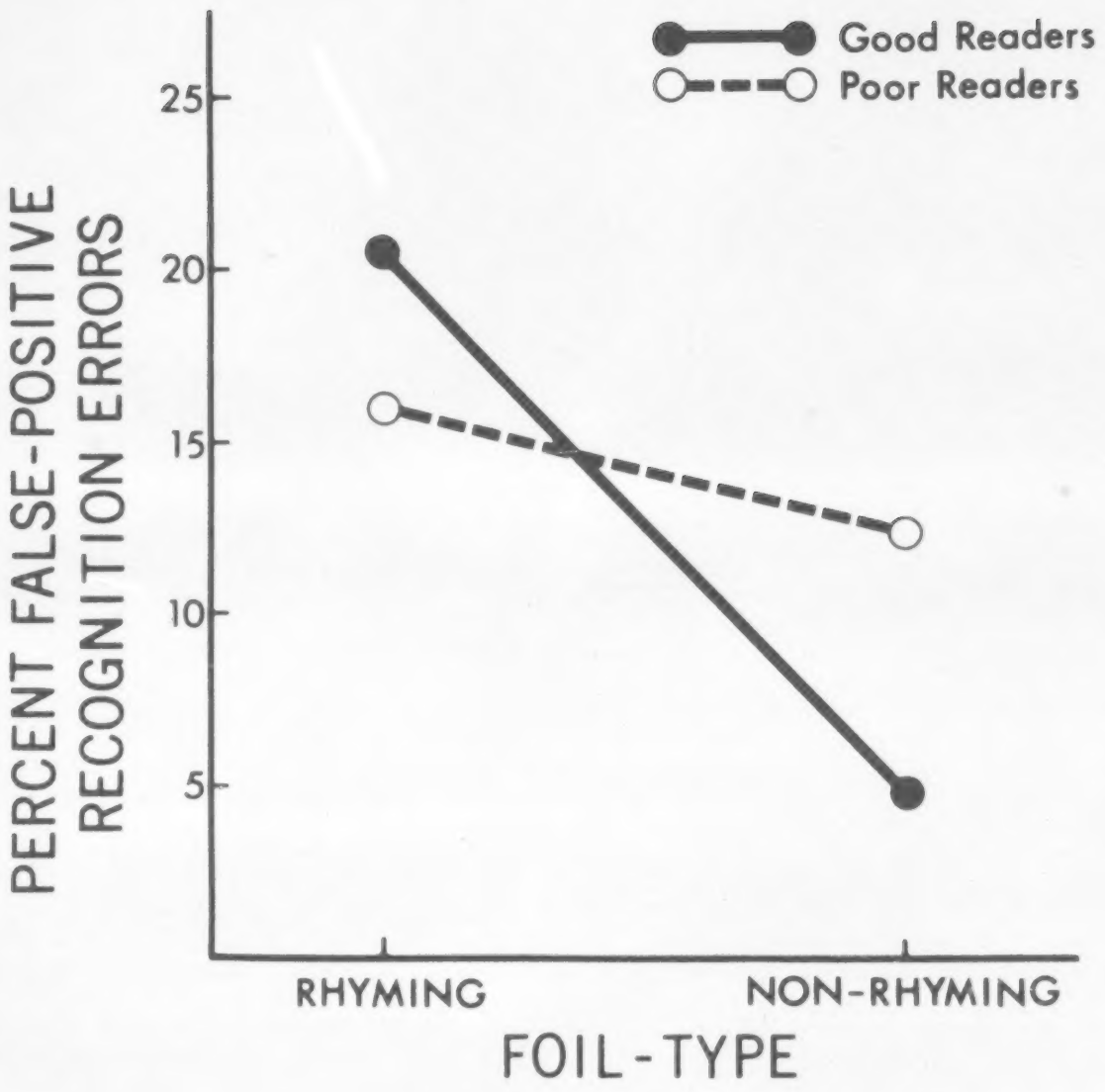


Figure 1: Percent false positive recognition errors as a function of reading ability and foil-type.

on the initial list when, in fact, it was a "new" word. The "false negative" error, which occurs when the child fails to recognize an "old word" as having appeared on the initial list, will also be considered.

False positive errors. The mean percentages of recognition errors for the two types of foils (rhyming and nonrhyming) were computed. For the good readers, the error rate was strikingly higher on the rhyming foils (20.4 percent) than on the nonrhyming foils (4.8 percent). In contrast, the poor readers showed little difference between the percentage of "false positive" errors made on the rhyming foils (16.0 percent) and the nonrhyming foils (12.4 percent). Because of the apparent heterogeneity of variance shown by the good readers on the nonrhyming foils relative to rhyming foils, a nonparametric statistic, the Mann-Whitney U-Test (Mann and Whitney, 1947) was used to assess the significance of the phonetic characteristics of the foils. For the good readers, the mean difference beween the mean recognition errors on the two foil categories was highly significant [$U(15,15) = 26$; $p < .002$], whereas for the poor readers the error difference between rhyming and nonrhyming foils was not significant [$U(14,14) = 80$; $p > .10$].

The interaction between reading ability and foil-type (Figure 1) was examined by comparing the difference between the error scores on the rhyming and nonrhyming foils for the two reading groups. The mean error difference was 15.5 percent for the good readers and 3.5 percent for the poor readers [$U(15,14) = 23.5; p < .002$]. These data strongly support the interpretation of the interaction between reading ability and responses to phonetic similarity that was offered by Liberman et al. (1977).

False negative errors. It is somewhat misleading to make a simple division of the old words into those with rhyming foils and those without a rhyming foil. On the recognition list, a word with a phonetically similar foil is indistinguishable from phonetically dissimilar old words until the appearance of its rhyming foil; only those old words that follow their rhyming foil on the recognition list can be said to differ from the nonrhyming old words. In comparing recognition judgments of rhyming and nonrhyming old words, it is reasonable to consider as "phonetically similar old words" only the words that appear after their rhyming foils; and consequently, all other repeated words must be viewed as nonrhyming old words. Using this criterion for categorizing old words, the frequency of "false negative" recognition errors for the good readers was 23.8 percent on the rhyming old words, and 28.8 percent on the nonrhyming old words. The comparable error rates for the poor readers were 18.8 percent and 19.6 percent respectively.

The pattern of false negative errors reflects a tendency on the part of the good readers to say that a word from the initial list was "old" when it followed its rhyming foil. Thus, for the good readers, words on the initial list that followed their rhyming foils on the recognition list more frequently evoked "yes" judgments than did words that lacked rhyming counterparts. The poor readers showed no such tendency. They made a nearly equal number of "yes" responses to phonetically similar and dissimilar words. Thus, the recognition judgments of repeated words reintorce the indications from the analysis of the false positive errors that good readers have a more persistent phonetic representation in short-term storage than do poor

readers.

Reading Errors

Table 2 shows the mean percentage of misread words by the good and poor readers on each of the four sets (phonetically-similar old words, phonetically-dissimilar old words, phonetically-similar foils, and phonetically-dissimilar foils) of words. As noted in the description of scoring procedures, recognition judgments of words that were misread on either list were not included in this tally. In addition, when a misread word rhymed with one of the foils on the recognition list, the recognition judgment on that foil was also excluded. As would be expected, the good readers made considerably fewer errors than the poor readers. In fact, 13 of the 15 good readers made no reading errors at all. The poor readers, on the other hand, misread an appreciable number of words. This is a matter for concern only if their errors are unequally distributed among the four sets of words. In that event, one could question the reliability of the differences in false positive recognition errors, the finding of major interest. However, from inspection of Table 2, it may be seen that roughly the same proportion of misreadings occurred on each of the four sets. This impression was substantiated by the results of a two-factor within-subjects analysis of variance in which phonetic similarity-dissimilarity was treated as one factor (P) and old and new (foil) words were treated as the other factor (R). Neither factor was significant [$F_P(1,13) < 1$; $F_R(1,13) < 1$]. It is apparent that the errors were indeed equally distributed among the four sets of words. Thus, the differences between the reading groups in the distribution of recognition errors on rhyming and nonrhyming foils cannot be attributed to a tendency on the part of the poor readers to make more errors in reading the words of some sets than of others.

TABLE 2: Reading errors as a function of opportunity for good and poor readers.

Reading Group		PS_f	PD_f	PS_o	PD_o
Good	Errors	6	1	4	2
n = 15	Opportunities	210	210	210	210
	Percent	2.9	0.5	1.9	1.0
Poor	Errors	27	30	30	34
n = 14	Opportunities	196	196	196	196
	Percent	13.8	15.3	15.3	17.3

PS_f - Phonetically Similar Foil
PD_f - Phonetically Dissimilar Foil
PS_o - Phonetically Similar Old Word
PD_o - Phonetically Dissimilar Old Word

DISCUSSION

In a recent study (Liberman et al., 1977), good beginning readers were found to be more affected than poor readers by the phonetic characteristics of visually-presented items in a recall task. We attributed this result to differences between the groups' abilities to employ phonetic representation. The possibility has been raised, however, that differences in rehearsal strategy may account for the finding. The major aim of the present experiment was to clarify the interpretation of the earlier study by using a task in which rehearsal was not a factor. For this purpose, a recognition memory paradigm was used instead of a recall task. The advantage of this procedure is that it does not alert the child to rehearse the target items, because he is not informed in advance that his memory of these items will be tested.

A secondary aim of the present experiment was to demonstrate the differential effects of phonetic similarity on good and poor readers in a task that employs words rather than arbitrary letter sequences, thus extending the earlier findings to a situation that more closely approximates an actual reading task.

The results are summarized in Figure 1: the good readers made fewer recognition errors on the nonrhyming foils relative to their performance on the rhyming foils; in contrast, the poor readers made roughly equal numbers of errors in recognition judgments on the two types of foils. The confirmation of the interaction between reading ability and phonetic similarity with this new task that minimizes possible rehearsal effects, suggests that the earlier findings cannot be attributed solely to differences in rehearsal strategy between good and poor readers. The data, therefore, tend to support the hypothesis that the two reading groups differ in their use of a phonetic representation.

It might be concluded, then, that poor readers have a specific difficulty in accessing a phonetic representation derived from script. There is reason to believe, however, that the poor readers' difficulties in making effective use of a phonetic representation are of a more general nature and not limited to recoding from script. The evidence comes from a study reported by Shankweiler and Liberman (1976) that was a sequel to the Liberman et al. (1977) visual recall experiment. The point of that study was to create an auditory analog of the earlier experiment, in which the letter strings would be presented on magnetic tape instead of tachistoscopically. Since phonetic coding is presumably unavoidable when speech is presented auditorily, both reading groups in the auditory experiment would thus be forced to code the incoming speech signal phonetically. If the poor readers' essential difficulty was specific to recoding visually presented script, the auditory version of the recall experiment should yield different results; the statistical interaction between reading ability and phonetic similarity, obtained in the previous study, should disappear. However, if the interaction remained, it would suggest that the phonetic recoding differences between good and poor readers are not specifically tied to the conversion from print to speech, but rather that the poor readers' deficit extends to heard speech as well as written language.

The results of these new experiments were nearly identical to those using visual recall. As before, the good readers showed significantly more phonetic interference than the poor readers. Thus, it may be concluded that the nature of the poor readers' deficit is related to the accessing and use of a phonetic representation, regardless of the source of the linguistic information. Further investigation of the circumstances that limit access to the phonetic representation is likely to contribute to an understanding of the sources of difficulty in learning to read.

REFERENCES

Baddeley, A. D. (1966) Short-term memory for word sequences as a function of acoustic and formal similarity. Q. J. Exper. Psychol. 18, 362-365.

Baddeley, A. D. (1968) How does acoustic similarity influence short-term memory? Q. J. Exper. Psychol. 20, 249-264.

Baddeley, A. D. (1970) Effects of acoustic and semantic similarity on short-term paired associate learning. Brit. J. Psychol. 61, 335-343.

Bender, L. (1957) Specific reading disability as a maturational lag. Bull. Orton Soc. 7, 9-18.

Benton, A. L. (1962) Dyslexia in relation to perception and directional sense. In Reading Disability: Progress and Research Needs in Dyslexia, ed. by J. Money. (Baltimore, Md.: Johns Hopkins Press).

Benton, A. L. (1975) Developmental dyslexia: Neurological aspects. In Advances in Neurology, vol. 7, ed. by W. J. Friedlander, (New York: Raven Press).

Bloomfield, L. (1942) Linguistics and reading. Elementary English 18, 125-130, 183-186.

Cheek, E. H. (1974) Cheek Master Word List. (Waco, Texas: Educational Achievement Corporation).

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-84.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Frostig, M. (1963) Visual perception in the brain-injured child. Am. J. Orthopsychiat. 33, 665-671.

Hammill, D. (1972) Training visual perceptual processes. J. Learn. Dis. 5, 39-46.

Helfgott, J. (1976) Phonemic segmentation and blending skills of kindergarten children: Implications for beginning reading acquisition. Contempor. Educ. Psychol. 1(2), 157-189.

Hintzman, D. L. (1967) Articulatory coding in short-term memory. J. Verbal Learn. Verbal Behav. 6, 312-316.

Hyde, T. S., and J. J. Jenkins. (1969) Differential effects of incidental tasks on the organization of recall of a test of highly associated words. J. Exp. Psychol. 82, 472-481.

Jastak, J., S. W. Bijou, and S. R. Jastak. (1965) Wide Range Achievement Test. (Wilmington, Delaware: Guidance Associates).

Kintsch, W., and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).

Liberman, I. Y. (1971) Basic research in speech and lateralization of language: Some implications for reading disability. Bull. Orton Soc. 21, 71-87.

Liberman, I. Y. (1973) Segmentation of the spoken word and reading acquisition. Bull. Orton Soc. 23, 65-77.

Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer. (1977) Phonetic segmentation and recoding in the beginning reader. In Toward a Psychology of Reading: The Proceedings of the CUNY Conferences, ed. by A. S. Reber and D. Scarborough. (Hillsdale, New Jersey: Lawrence Erlbaum Assoc.).

Mann, H. B. and D. R. Whitney. (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18, 50-60.

Mattingly, I. G. (1972) Reading, the linguistic process and linguistic awareness. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Money, J. (1962) Reading Disability: Progress and Research Needs in Dyslexia. (Baltimore, Md.: Johns Hopkins Press).

Rozin, P. and L. R. Gleitman. (1977) The structure and acquisition of reading. In Toward a Psychology of Reading: The Proceedings of the CUNY Conferences, ed. by A. S. Reber and D. Scarborough. (Hillsdale, New Jersey: Lawrence Erlbaum Assoc.).

Savin, H. B. (1972) What the child knows about speech when he starts to learn to read. In Language by Eye and by Ear: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. and I. Y. Liberman. (1976) Exploring the relations between reading and speech. Paper presented at the Conference on the Neuropsychology of Learning Disabilities, Korsor, Denmark, June 1975. Published in The Neuropsychology of Learning Disorders: Theoretical Approaches, ed. by R. M. Knights and D. K. Bakker. (Baltimore, Md.; University Park Press).

Silver, A. and R. Hagin. (1960) Specific reading disability, delineation of the syndrome and relationship to cerebral dominance. Comp. Psychiat. 1, 126-134.

Vellutino, F. R., J. A. Steger, and G. Kandel. (1972) Reading disability: An investigation of the perceptual deficit hypothesis. Cortex 8, 106-118.

Vellutino, F. R., R. M. Pruzek, J. A. Steger, and U. Meshoulam. (1973) Immediate visual recall in poor and normal readers as a function of orthographic-linguistic familiarity. Cortex 9, 368-384.

Vellutino, F. R., J. A. Steger, C. J. Harding, and F. Phillips. (1975) Verbal vs. non-verbal paired-associates learning in poor and normal readers. Neuropsychologia 13, 75-82.

Vernon, M. D. (1960) Backwardness in Reading. (Cambridge: Cambridge University Press).

Wechsler, D. (1974) Wechsler Intelligence Scale for Children - Revised. (New York: The Psychological Corporation).

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Am. 39, 588-598.

Interactive Experiments with a Digital Pattern Playback*

Patrick W. Nye, Franklin S. Cooper, and Paul Mermelstein

ABSTRACT

Among the most useful tools for speech research have been those that enable spectrograms to be compared with one another, that provide ways of modifying speech data and that permit the user to listen to the modified speech signal. This paper reports an experiment in which such an interactive research tool--a Digital Pattern Playback (DPP)--was used to evaluate a spectrum-matching and dictionary-search technique for speech recognition. The DPP was used to display spectrograms of "unknown" sentences. An analyst divided these sentences into segments of word-length and listed their important acoustic features. Using these features, an interrogation program examined a feature-based spectrographic dictionary and recovered all the words having features that matched each unknown segment. When necessary, additional features were assigned to narrow the search. The reference spectrograms retrieved from the dictionary were compared, one at a time, with the spectrograms of the unknown sentence, and the best match was selected for each unknown segment. In general, the performance of the human analysts was found to be quite low, since only 26 percent of the words contained in the sentences were matched correctly. The paper concludes with a discussion of the factors governing human and machine performance on spectrogram matching.

INTRODUCTION

This paper describes results obtained from a speech analysis experiment that explored methods for organizing the information required for automatic speech recognition. The experiment required that the analysis operations be performed by two human subjects who worked from visual displays. These analysts studied the spectrogram, waveform, and amplitude functions of an unknown sentence and divided the sentence into word-length segments. Having listed the most salient features of each segment, the analysts then sought a set of matching reference words that were retrieved automatically from a feature-labeled dictionary. The identities of the reference words were not known to either of the analysts whose data are reported in this paper. Thus, syntactic and semantic considerations did not play a direct part in the selection of suitable matches.

---

*This paper was presented in part at the 90th meeting of the Acoustical Society of America, San Francisco, Calif. November 3-7, 1975.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

Ingemann and Mermelstein (1975) have reported the results of some similar experiments that were carried out with conventional paper spectrograms. Their experience showed that the clerical problems became serious when subjects were required to work with reference libraries as large as 100 words. The present work represented a continuation of those experiments but avoided the inconvenience of handling volumes of paper by using a computer-based display system.

THE DISPLAY SYSTEM

The speech signals were displayed by an interactive research tool--called the Digital Pattern Playback (DPP)--which has been built around a PDP 11/45 and GT40 computer system (Nye, Reiss, Cooper, McGuire, Mermelstein, and Montlick, 1975). The system organization is sketched in Figure 1. The PDP 11/45 runs a general-purpose operating system allowing multiprogram access from several terminals. The GT40 supports the display functions. The analyst, seated at the keyboard, can selectively access the PDP 11/45 or the GT40. Using this facility, he may display two spectrograms lying one above the other on the same screen--each representing 1.6 secs of speech (see Figure 2). The lower spectrogram display field is usually occupied by a reference item that has been selected from the dictionary and installed there for direct comparison with the unknown. A cursor, controlled by a knob, can be moved to any point along the time axis of the upper, unknown spectrogram and the cross-section at that point can be displayed. A similar cross-section facility is also available for the lower spectrogram. In addition, the user has the freedom to examine waveform plots for the unknown at points indicated by the cursor, and to examine the intensity and fundamental frequency functions of selected segments of speech data. Other facilities include provisions for manipulating speech spectra and hearing the results through a channel vocoder. The system forms a general speech analysis-synthesis facility, only a few of whose capabilities were employed in the experiments described here.

ORGANIZATION OF THE RETRIEVAL PROGRAM

Each of the reference spectrograms consisted of a candidate word presented in the sentence frame "Please say again." These spectrograms made up a lexicon of 100 reference items of which 20 had both stressed and unstressed forms represented, giving a grand total of 120 entries. The items were stored on a disk in such a way that they could be selectively retrieved by means of a specially designed program that also collected data on each analyst's decisions and analysis procedures. A general model of this process is given in Figure 3.

Before commencing the experiment, the two analysts were each asked to select a personal set of up to 16 descriptive features that were considered to be useful in correctly selecting matching words from the lexicon. Each analyst then used his chosen features to label each member of the reference list. Any one of three discrete values could be assigned to each feature; either present, absent or unspecified.

The retrieval program listed the features that an analyst found in a word-segment of the unknown sentence and used this list (or feature vector)

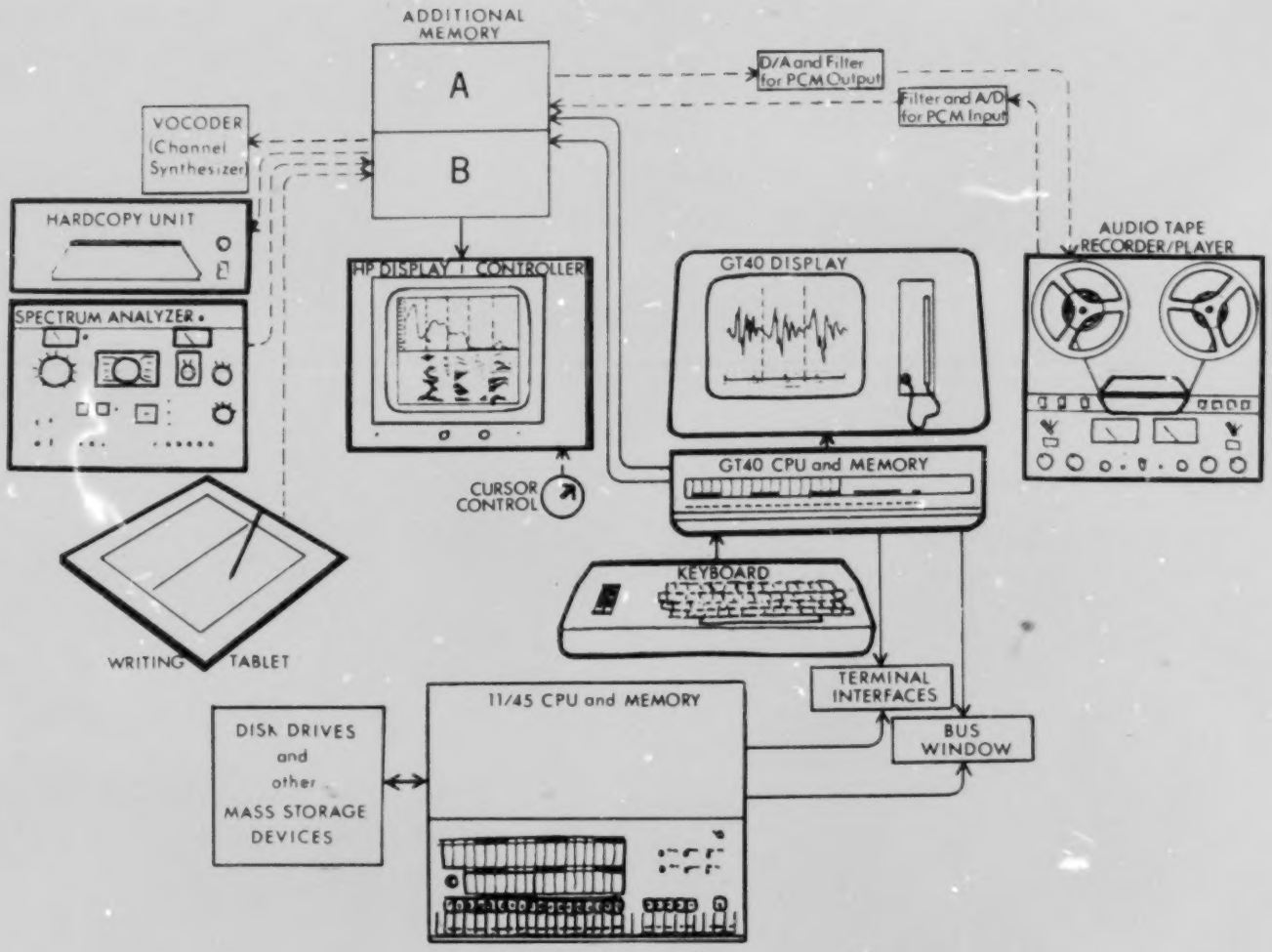


FIGURE 1

Figure 1: An overall view of the DPP showing its principal components. The physical components actually used by the analysts were the HP display and cursor, the GT40 display and the keyboard.

FIGURE 2

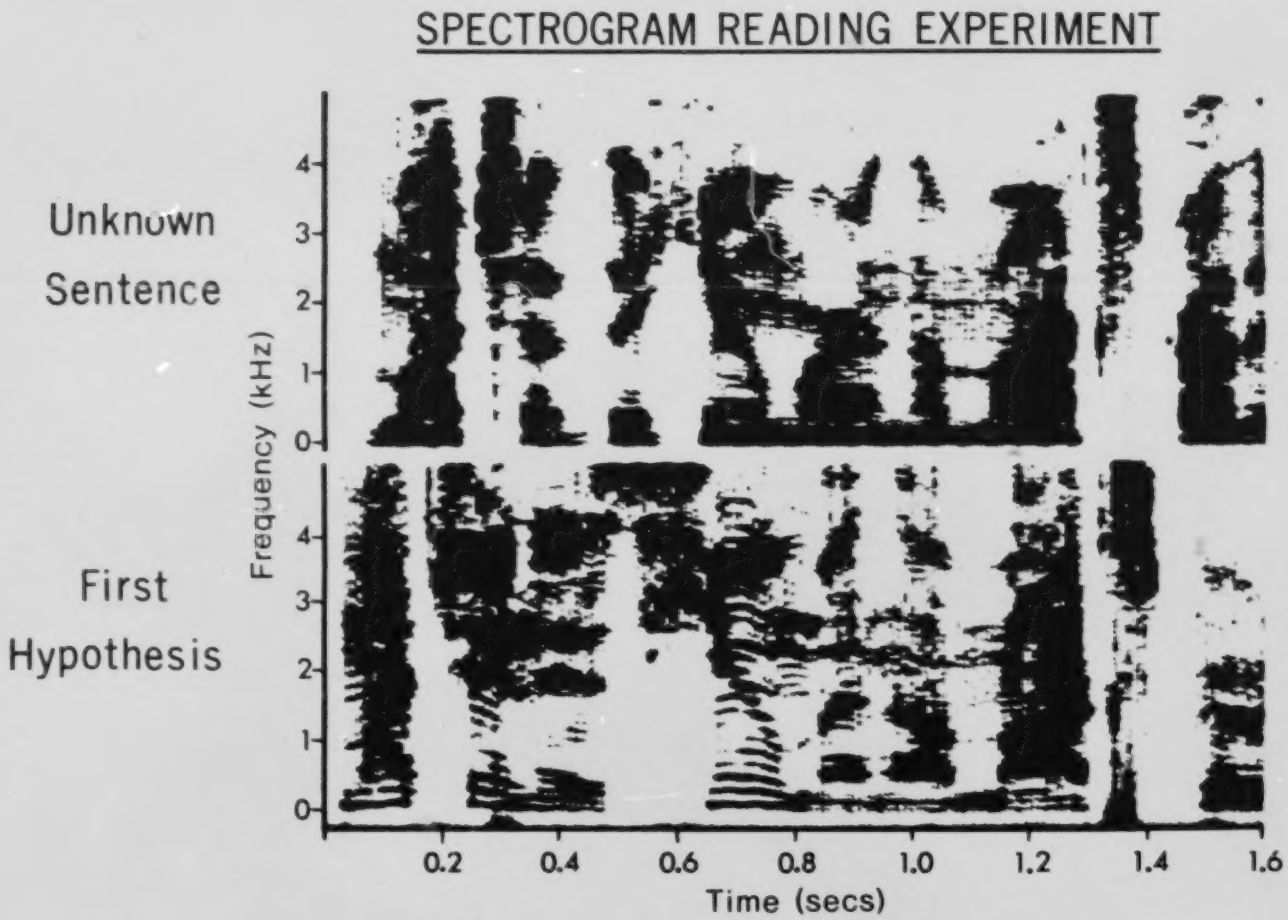


Figure 2: A copy of the computer-displayed spectrogram of a portion of the unknown sentence (top) and the hypothesis obtained from the first analyst after pass 1 (bottom).

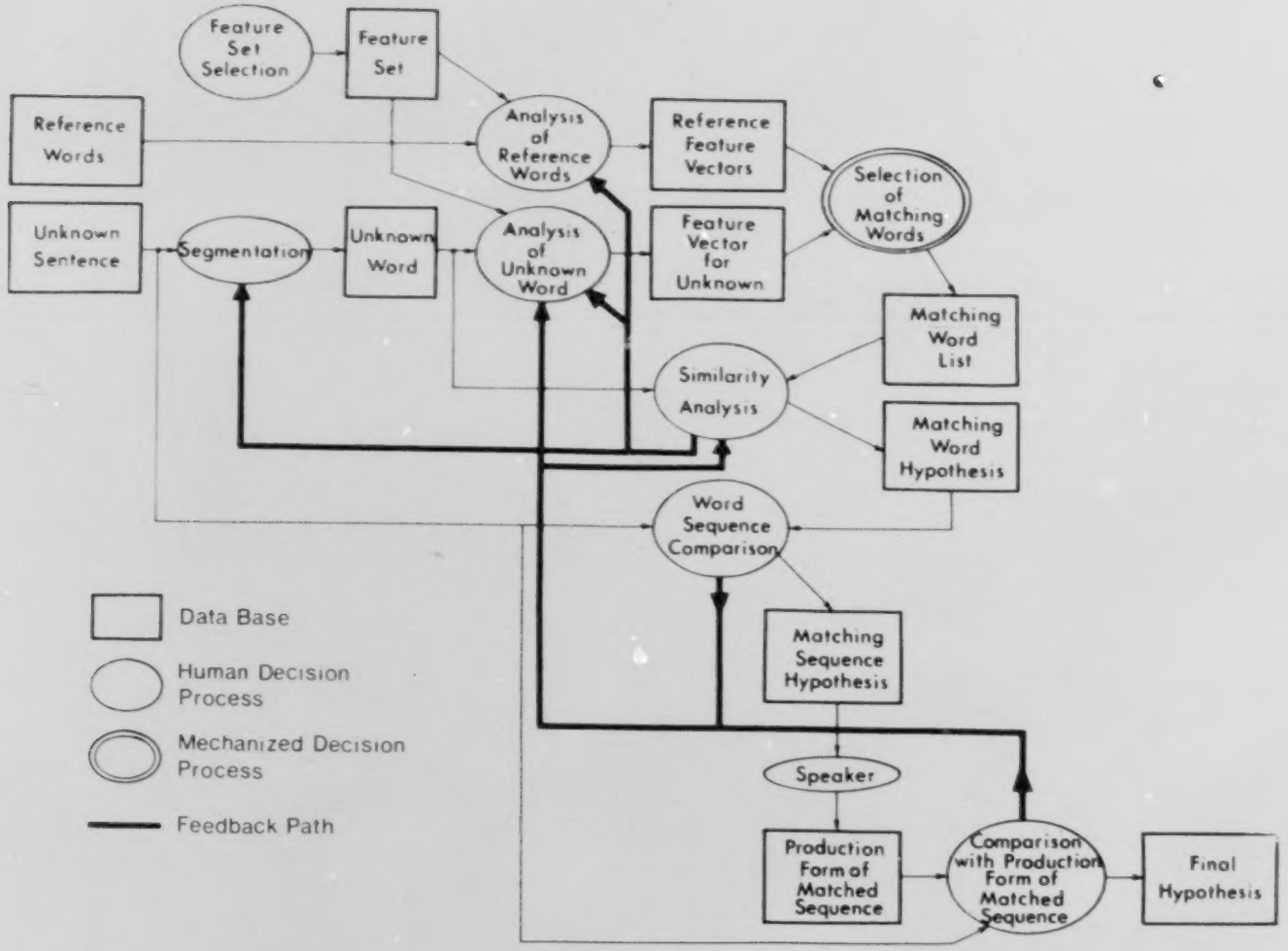


FIGURE 3

Figure 3: Model of the information pathways available to the analysts who formulated an hypothesized spectrogram from a sequence of reference-word units.

to extract a subset of the reference lexicon that shared the same features. An additional property of the program was that "unspecified" feature values, in reference words, matched both "present" and "absent" assignments of those features in the unknown segment.

While the matching program was under way, an analyst could specify additional feature information about the unknown by changing values or specifying previously unspecified values. Alternatively, he could relax feature assignments by increasing the number of unspecified features and thereby increase the size of the matching word list. The number of reference spectrograms that matched any specified feature-vector could be rapidly determined. In the event that too many reference items matched the specified features, the analyst was allowed to revalue features in the reference list to achieve greater precision. When the number of retrieved items fell to a sufficiently low level, the analyst could scan through them one by one, each time displaying the potential match above the unknown. In order to make a unique selection, he could then invoke additional information not included in the previous feature assignment; for example, expected formant shifts from the reference form to fit the apparent context of the unknown. If none of the retrieved items matched sufficiently well, the feature assignment was then modified to select a new list of matching words.

The analyst could also display a series of potential word matches in an appropriate order, side by side, and judge whether coarticulation effects could account for the remaining discrepancies between the reference words and the unknown. After the analyst had arrived at a hypothesized reference-word sequence that satisfied his criteria, the sequence of items was given to the original speaker to be spoken in the same tone of voice and with the same intonation pattern used in the original unknown sentence. This production form of the matched sequence was then added to the data base for the analysts' examination. At this point, new reference words could be substituted where the analyst noted that a mismatch with the unknown sentence had occurred.

The record-keeping section of the retrieval program noted the number of searches of the reference library that were made by both the analysts and all of the reference words that they examined. This record allowed the authors to trace the significant information feedback paths in the system--those that resulted in new searches of the reference library with differing feature-vectors. These feedback paths are noted in Figure 3. The extent to which lexical information can modify an analyst's segmentation and feature assignment was not surprising. In fact, through this attempt to model explicitly the information flow among the various subtasks of the analysis process, we have uncovered a structure similar to the model for speech recognition proposed by Fant (1970) nearly 7 years ago.

EXPERIMENTAL OBSERVATIONS

Both analysts found the 16 assignable features to be insufficient and would have used a larger number, had there been provision to do so. However, even the assignment of sixteen features to 120 reference items was very time-consuming. In order not to impose any prior feature organization on our analysts, all features were considered equally important in establishing a

match. The analysts were frustrated by the necessity to explicitly mark the absence of many features--a requirement imposed by the single-level feature organization. Use of a multilevel or hierarchic feature organization necessitating the selection of secondary features only if they were appropriate in the light of specific assignments of the primary features, would have overcome this difficulty.

Both analysts found little difficulty establishing reference-word matches to the prominent words of the unknown sequence. In fact, they were surprised to discover how little information (possibly only 3 or 4 features) sufficed for the retrieval of no more than 6 matching items. More severe difficulties were encountered in attempting to find the matches for the less prominent words or syllables. Here the analysts did not trust their feature assignments--an indication of the difficulty that they encountered in making those assignments in the first place. One analyst resorted to an exhaustive scanning of the list of unstressed reference items. The other compared pairs of stressed and unstressed reference items to infer which features could be expected to be harder to detect under reduced stress. He then relaxed the feature assignment for the corresponding unstressed items.

The second analyst attempted to overcome the word segmentation problem by selecting prominent syllables around which to organize a retrieval attempt. The ability to look at variations in the spectrum envelope as the cursor swept through successive time intervals of the spectrograms proved to be quite helpful in selecting the most prominent syllable of a sequence. Organizing the retrieval strategy around prominent syllables permitted the rapid examination of alternative hypotheses. For example, the first hypothesis might be a monosyllabic stressed word, the second a bisyllabic word with an additional unstressed syllable. Information about additional consonantal segments could be added to the feature vector used for retrieval until the number of retrieved items was small enough to be individually scanned. Even though only a few salient features located near the prominent vowel were assigned, the retrieval process frequently resulted in an obvious match to a much longer segment of the unknown.

The features describing vowel color were not found very useful by either analyst. There are two reasons that may account for this finding. First, contextual influences on the vowel formant-frequencies of both the reference word and the unknown word-segment made reliable feature assignment difficult. Second, very few of the reference items differed by vowel color alone. Thus, the specification of vowel color features did not significantly reduce the number of retrieved matches in contrast to leaving them unspecified.

The one analyst who attempted to make use of segment duration in his feature assignment found it to be useful only in extreme cases. For the most part, the segmental durations of unknown words varied considerably as a function of stress, syntactic role and position in the sentence, making small durational differences ineffective for discrimination purposes.

Results

The average proportion of words that the two analysts succeeded in correctly matching was only 26 percent, and this figure did not increase after one cycle of feedback. Although one error was corrected, an additional error was introduced in the words hypothesized on the second attempt. The overall word-matching performance was thus significantly lower for the machine-assisted word-matching experiment than for the similar experiment conducted with conventional spectrograms by Ingemann and Mermelstein (1975). There are several possible reasons for this deterioration in matching performance. The relative unfamiliarity of the display--in particular the way acoustic features seen on the DPP are affected by the limited time resolution of the display--may have been one factor. More importantly, perhaps, the sentence in the current experiment was longer (21 words vs. 16 words) and somewhat more complex. The lexicon used in the DPP experiment intentionally included more words that had close phonetic similarities to the unknown words of the sentence.

The word-identification scores are broken down by analyst, stress, and number of syllables in Table 1. While 52 percent of the words that contained at least one stressed syllable were correctly identified overall, practically all of the matches with unstressed words were incorrect. Overall performance on multisyllabic words was somewhat higher than on monosyllabic words. Here the relative performance of the subjects differed significantly. The analyst who used the strategy that focused on prominent syllables did better on monosyllabic words but worse on multisyllabic words. The strategy led to frequent errors on the unstressed syllable of a multisyllabic word--particularly when phonetically similar words were included in the lexicon. Substitutions in the unstressed syllables of those words were quite frequent. Examples of such substitutions are "immunity" for "community", "human" for "humor", "arrive" for "derived", and "salt" for "assault."

TABLE 1: Percent correctly identified words.

		Pass 1		Pass 2	
	Tokens	Analyst 1	Analyst 2	Analyst 1	Analyst 2
Monosyllabic words	15	33	20	27	20
Multisyllabic words	6	17	33	17	50
Unstressed words	11	9	0	0	0
Stressed words	10	50	50	50	60
All words	21	29	23	23	29
		Percent correctly identified syllables			
All syllables	30	43	47	43	47

CONCLUSIONS

The single most important observation to emerge from the results is the poor performance of the analysts on unstressed words. A reference token, whether spoken in a stressed form or in a different unstressed environment, does not provide sufficient information to enable the analyst to effect a match. Perhaps a larger number of reference tokens taken from a variety of contexts in which the word may occur might be useful, since it is evident that analysts are usually unable to predict the transformations that the acoustic features of words can undergo if they are uttered in phonetically different contexts. Analysts generally judge similarity in terms of common features between the unknown and reference tokens. They do not pay particular attention to the variability of those features and thus do not differentiate among the features according to their reliability in establishing matches. It seems likely that intensive learning sessions on the variability of acoustic features are required before improved word matching results can be obtained.

The lack of any significant improvement following feedback of the hypothesized words spoken as a sentence is probably due to the fact that the overall performance was initially too low (that is, the initial hypothesis was offered with such a low level of confidence that it contributed as much to the analyst's uncertainty as it did to his knowledge). It appears to be that a higher minimum performance must be reached before the information supplied by feedback can be usefully absorbed. If an unknown word is embedded in the correct context, its appearance is likely to be quite similar to its form in the unknown sentence. However, if the context is incorrect as well, a new production of the reference form is obtained that may not be any more similar to the unknown than it was to the original.

Let us now consider the prospects for implementing an entire feature assignment and word-matching procedure in algorithmic form for execution by a machine. The selection of matching words on the basis of assigned feature values is clearly the easiest procedure to implement, and, in fact, this has already been successfully carried out. Heuristics are available for the assignment of values to most acoustic features and, therefore, we can expect that this analysis procedure can be implemented at a cost that increases roughly linearly with the number of features used. We anticipate more difficulty, however, with the process labeled "similarity". We are not, as yet, able to quantify a general similarity metric that assigns perceptually appropriate weights to specific differences. Events of short duration, such as bursts, may contribute a great deal to measures of similarity, whereas differences in events of longer duration, such as shifts in formant frequencies in vocalic intervals, may be of less significance.

It is possible that the comparison of word-sequences might be implemented with the aid of a speech synthesis program; however, it appears that finding an appropriate metric of similarity is the most difficult problem. Given any general difference measure, we do not yet know how to separate differences between speakers from differences between words, and until we can learn what the important distinctions are that we must look for, word identification through spectrum matching by a human analyst, or by a machine, will not be a practical art.

REFERENCES

Ingemann, F. and P. Mermelstein. (1975) Speech recognition through spectrogram matching. J. Acoust. Soc. Am. 57, 253-255. [Also Haskins Laboratories Status Report on Speech Research SR-39/40, 53-65, (1974)].

Nye, P. W., L. J. Reiss, F. S. Cooper, R. M. McGuire, P. Mermelstein, and T. Montlick. (1975) A digital pattern playback for the analysis and manipulation of speech signals. Haskins Laboratories Status Report on Speech Research SR-44, 95-107.

Fant, G. (1970) Automatic recognition and speech research.. Speech Transmission Laboratory Quarterly Progress and Status Report 1/1970 (Stockholm, Sweden: Royal Institute of Technology). [Also in G. Fant, Speech Sounds and Features, (Cambridge, Mass.: MIT Press) 1973.]

The Function of Strap Muscles in Speech: Pitch Lowering or Jaw Opening?*

James E. Atkinson† and Donna Erickson

ABSTRACT

This paper reports on one aspect of a continuing study to determine the physiological correlates of the changes in fundamental voice frequency (F_0). Several electromyographic (EMG) studies with speech have reported an association of strap muscle activity, particularly the sternohyoid, with low F_0 and some of these studies suggest that the sternohyoid is actively involved in lowering F_0. It has also been suggested, however, that the sternohyoid is involved with jaw opening, and that the reported pitch-lowering effects may actually be the result of jaw opening. To investigate this question an EMG experiment was conducted on one speaker of American English under normal and clenched jaw conditions. The normal utterances were of the form "Bev loves Bob" with emphasis on the various words. The clenched jaw data were obtained while the subject held his jaw fixed by biting on a tongue depressor and intoned the corresponding intonation patterns with a fixed vowel carrier /a/. The results indicate that the strap muscle activity for the normal utterances is very similar to the activity for the same intonation pattern with the jaw clenched. Strap muscle activity thus seems to be more closely related to pitch effects than to jaw-opening effects.

This paper reports on one aspect of a continuing study to determine the physiological correlates of changes in fundamental voice frequency (F_0). Specifically, we investigate the sternohyoid muscle, one of several extrinsic laryngeal muscles, and its role in controlling F_0. Several electromyographic (EMG) studies with speech have reported an association of strap muscle activity, particularly the sternohyoid with low F_0, and some of these studies suggest that the sternohyoid is actively involved in lowering F_0 (Faaborg-Andersen, 1965; Ohala, 1970; Ohala and Hirose, 1970; Atkinson, 1973; Collier, 1975; Erickson, 1975). It has also been suggested, however, and there has been some supportive data, that the sternohyoid plays a role in some

---

*A version of this paper was presented at the 92nd meeting of the Acoustical Society of America, San Diego, California, November, 1976.

†Special Projects Department, Naval Underwater Systems Center, New London, Connecticut.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

articulatory gestures involving jaw opening and that the reported pitch lowering effects may actually be the result of jaw opening (Ohala, 1972; Ohala and Hirose, 1970; Harris, 1971).

Figure 1 gives a simplified schematic representation of the relevant anatomy. First, the thyroid cartilage and larynx as a whole are suspended from the hyoid bone by the strap muscles (hence their name). Clearly, contraction of these muscles can affect the thyroid cartilage in either the front-back or in the vertical direction. Any such movement could change the length and tension of the vocal cords and hence their rate of vibration (F_0). The exact mechanism involved is still not clear, although several possible explanations have been suggested.

The figure shows only one of the supra-hyoid muscles, for simplicity, the digastric muscle, although there are other muscles (such as geniohyoid and mylohyoid) in this group. Both the strap muscles supporting the larynx and the jaw opening muscles attach to the hyoid bone. As seen in Figure 1, contraction of the digastric creates a force that pulls the hyoid bone upward. To allow jaw opening, there must be an opposing downward force to stabilize the hyoid and give the jaw opening force something to pull against. Thus, it has been suggested that the sternohyoid and/or other strap muscles contract to supply this force and allow jaw opening.

To investigate this question, an EMG experiment was conducted on one speaker of American English under normal and clenched jaw conditions. The normal utterances were: "Bev loves Bob," with emphasis on various words. The clenched jaw data were obtained while the subject held his jaw fixed by biting on a tongue depressor and intoned the corresponding pitch patterns with a fixed vowel carrier /a/. An example is "BEV loves Bob" with the corresponding clenched jaw form "AH hah hah." A direct comparison of sternohyoid activity for the same pitch pattern with and without jaw opening effects was obtained.

Table 1 lists the utterances used.

TABLE 1: Test utterances.

NORMAL	CLENCHED JAW
BEV loves Bob.	AH hah hah.
Bev LOVES Bob.	ah HAH hah.
Bev loves BOB.	ah hah HAH.

EMG data were obtained from the sternohyoid muscle using hooked wire electrodes and then recorded for processing and analysis using the Haskins Laboratories EMG facility.

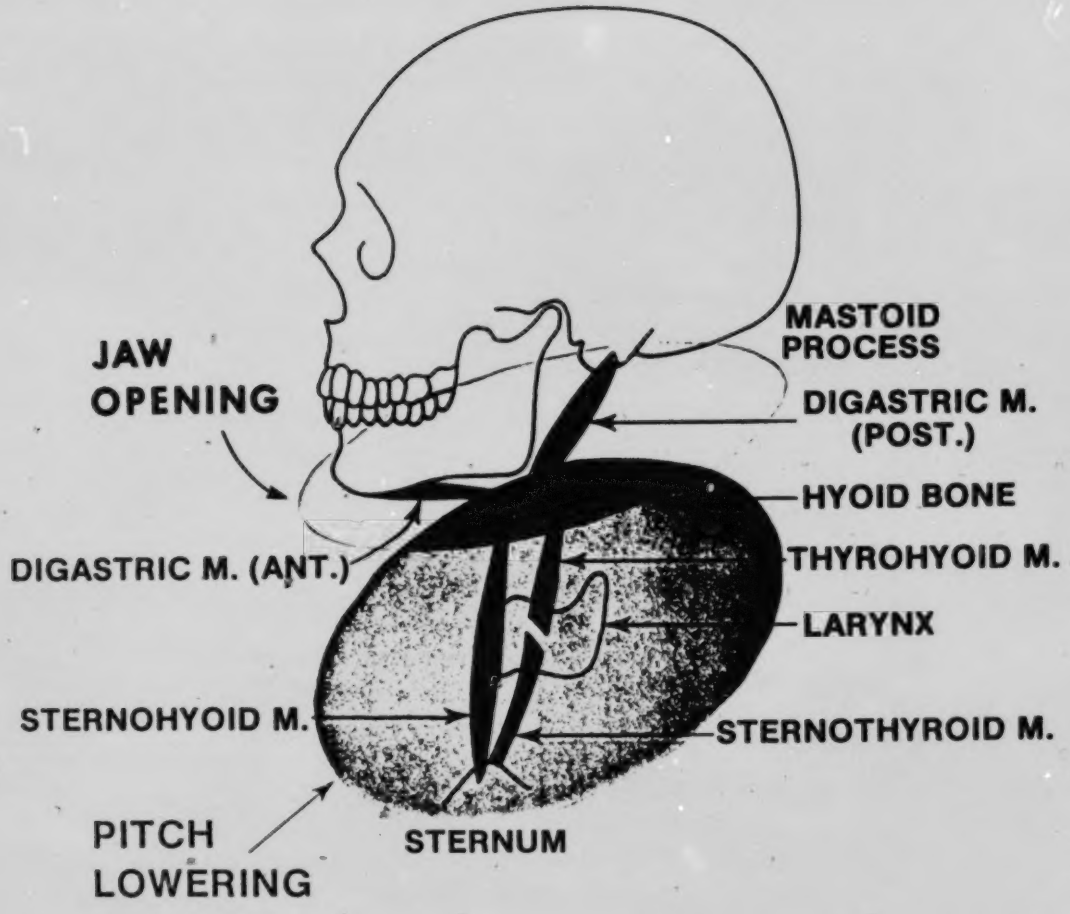


Figure 1: Simplified schematic representation of the muscles involved in pitch lowering and jaw opening.

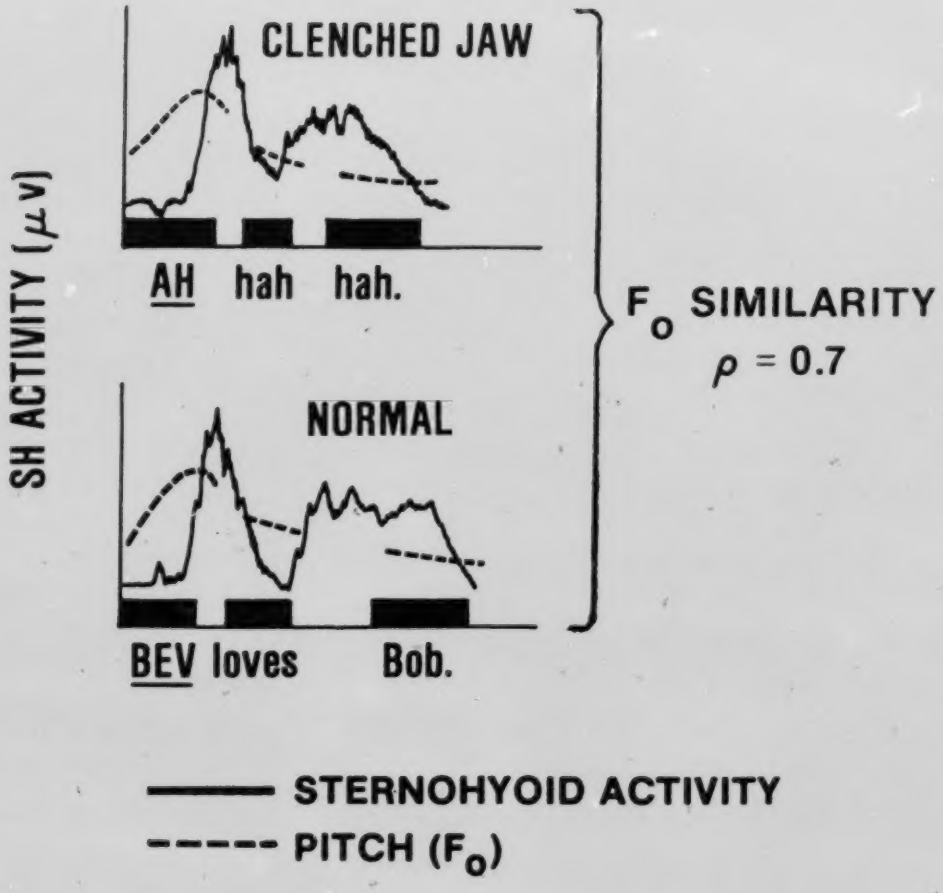


Figure 2: Comparison of sternonyoid muscle activity for normal and clenched jaw versions of an utterance having the same intonation pattern.

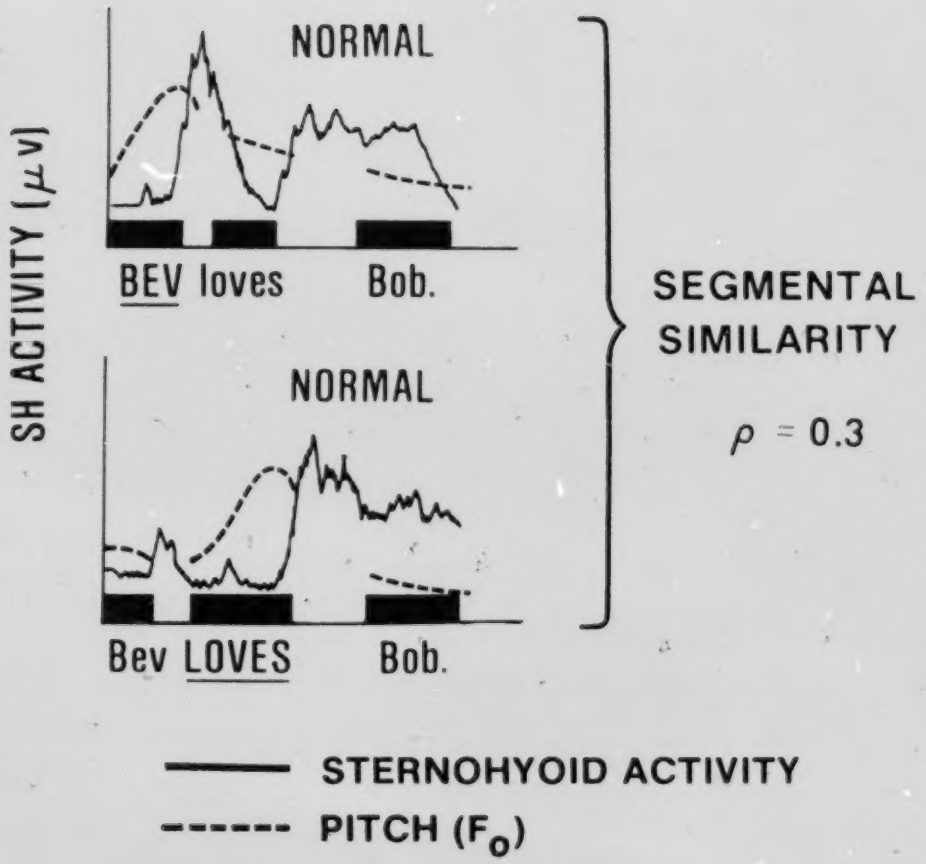


Figure 3: Comparison of sternohyoid muscle activity for two utterances having the same segmental phonemes but different intonation patterns.

The major results are given in Figure 2, which shows sternohyoid activity for the normal utterance "BEV loves Bob," and for the clenched jaw version with the same F_0 pattern "AH hah hah." In comparing these two utterances we are, in effect, holding pitch constant, and any differences in muscle activity must be a result of articulatory and jaw opening effects. Although there are some timing differences, as seen in this figure, it is quite clear that the sternohyoid activity is very similar for both the normal and clenched jaw versions. In fact, even with the timing differences the waveforms have a correlation coefficient of 0.7. Thus, no noticeable jaw opening effect is shown.

In Figure 3 we compare sternohyoid activity for the normal utterances "BEV loves Bob" and "Bev LOVES Bob." Here we effectively have the same segmental and jaw opening effects but very different F_0 patterns. Any differences in muscle activity thus would seem to be caused by pitch differences.

Clearly, the muscle activity is less similar here than in Figure 2 (the correlation coefficient is only 0.3). Thus, a clear pitch effect is seen.

To summarize, utterances having the same pitch pattern regardless of articulatory differences show very similar sternohyoid activity. Utterances having the same articulatory and jaw opening gestures (but different pitch patterns) show very different sternohyoid activity. We conclude, therefore, that sternohyoid activity is more closely related to pitch effects than to jaw opening effects, at least in this speaker. We are presently extending the study to other speakers in order to test the generality of these conclusions.

REFERENCES

Atkinson, J. E. (1973) Aspects of intonation in speech: Implications from an experimental study of fundamental frequency. (Unpublished Ph.D. Dissertation, University of Connecticut).

Collier, R. (1975) Physiological correlates of intonation patterns. J. Acoust. Soc. Am. 58, 249-255.

Erickson, D. M. (1976) A physiological analysis of the tones of Thai. (Unpublished Ph.D. Dissertation, University of Connecticut).

Faaborg-Andersen, K. (1965) Electromyography of Laryngeal Muscles in Humans: Techniques and Results. (Basel: S. Karger).

Harris, K. S. (1971) Action of the extrinsic musculature in the control of tongue position: preliminary report. Haskins Laboratories Status Report on Speech Research SR-25/26, 87-96.

Ohala, J. J. (1970) Aspects of the control and production of speech. Working Papers in Phonetics (University of California at Los Angeles) 15.

Ohala, J. J. (1972) How is pitch lowered? J. Acoust. Soc. Am. 52, 124(A).

Ohala, J. and H. Hirose. (1970) The function of the sternohyoid muscle in speech. (Research Institute of Logopedics and Phoniatrics, University of Tokyo, Annual Bulletin) 4, 41-44.

The Geniohyoid and the Role of the Strap Muscles*

Donna Erickson, Mark Liberman† and Seiji Niimi

ABSTRACT

Many investigators have noted a relationship between strap muscle activity and pitch lowering, but there does not seem to be any single generally accepted theory to account for this connection. The particular effect of strap muscle contraction will depend in part on what other forces are acting on the hyoid bone; therefore, in the context of a general EMG investigation of English intonation, we recorded from a suprahyoidal muscle, the geniohyoid, as well as the strap muscles (sternohyoid, sternothyroid, and thyrohyoid) and the cricothyroid. In our data, the three strap muscles show nearly identical patterns of activity; as a group, their activity shows a strong negative correlation with the activity of the geniohyoid and the cricothyroid. Examination of the relationship of these muscles' activity to F_0 levels showed the cricothyroid and geniohyoid to have a positive relation to F_0, and the sternohyoid (selected as a representative strap muscle) to have a slightly negative relation to F_0. These findings are related to the development of a possible model for the relative motion of the larynx during pitch changes.

EXPERIMENT

It is known that the strap muscles [sternohyoid (SH), sternothyroid (ST) and thyrohyoid (TH)] are active during low and falling F_0 (Ohala, 1970; Ohala and Hirose, 1970; Atkinson, 1973; Collier, 1975; Erickson, 1976). The suprahyoidal muscles have not previously been investigated with respect to their role in F_0 control in speech; yet, the anatomical arrangement of extrinsic laryngeal musculature is such that an effect of the strap muscles with respect to F_0 will certainly depend in part on suprahyoidal forces as one can see by referring to Figure 1. In view of these considerations, we examined the EMG activity from a representative suprahyoidal muscle, the geniohyoid (GH), as well as from the three strap muscles (the SH, ST, and TH), and the cricothyroid (CT) in the context of a larger EMG experiment on English intonation.

We will present data from this experiment that bears on the following two questions:

---

*The paper was presented at the 92nd meeting of the Acoustical Society of America, San Diego, California, 15-29 November 1976.

†Bell Laboratories, Murray Hill, N.J.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

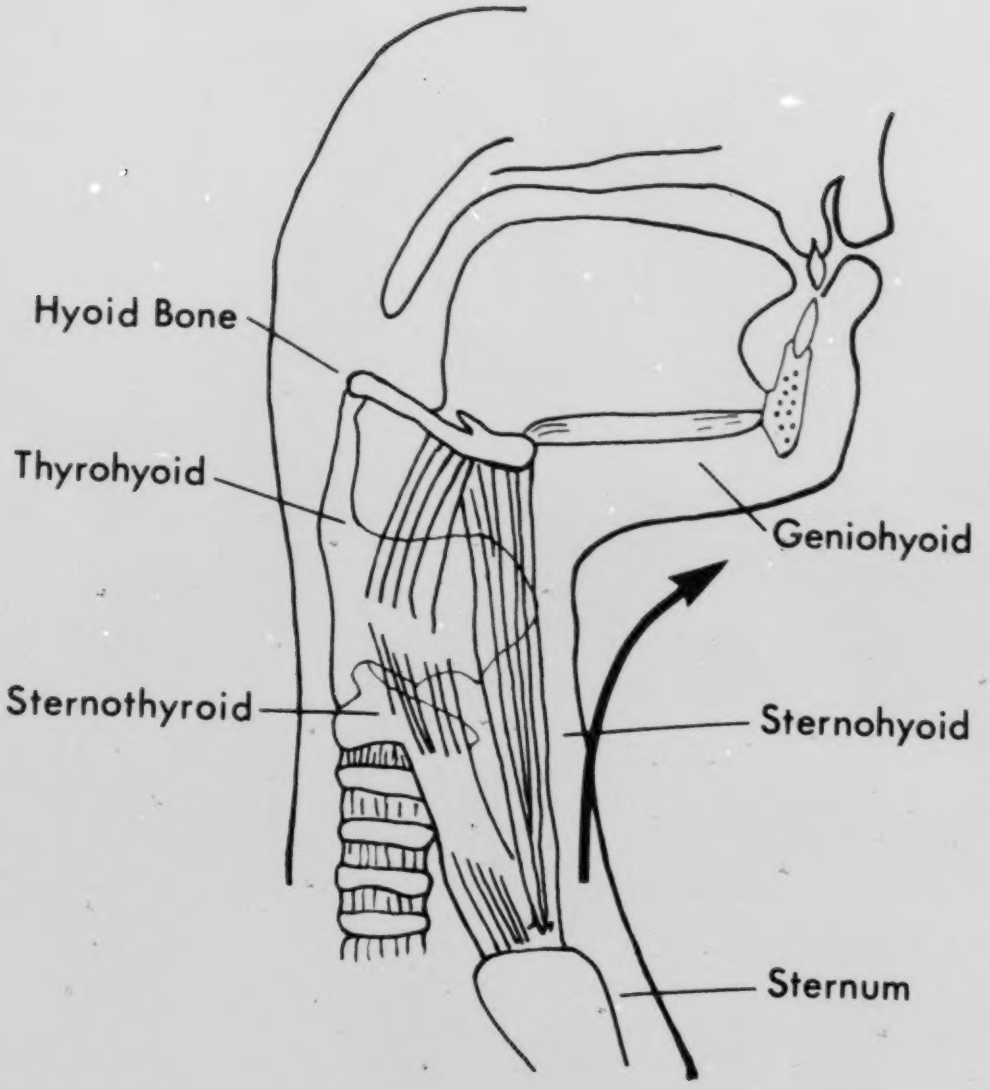


Figure 1: Extrinsic laryngeal muscles. It is hypothesized that the larynx tends to move as a whole in an arc, as shown by the arrow.

1. What overall relationship do the activities of the strap muscles, cricothyroid, and geniohyoid bear to each other?

2. What relationship do the activities of these muscles bear to F_0 levels?

Eight sentences, from 9 to 14 syllables long, with various stressings and intonational patterns, repeated eight times, were examined. We have only recorded a single speaker thus far. The quantitative data we will present in this paper is drawn from a subset of the total, but the results are qualitatively valid for the entire experimental run.

Results

Relationships among the muscles. Pearson product-moment correlation coefficients were calculated for the various muscles using the Haskins Laboratories computer-implemented correlation program, and the results can be seen in Figure 2. The activity of the three strap muscles positively correlates with each other, and negatively correlates with the geniohyoid and cricothyroid. The activity of the geniohyoid and cricothyroid positively correlates with each other. Although quantification of the correlation of intrastrap muscle activity in terms of correlation coefficients has not been presented heretofore in the literature, the data here agree with the findings of Erickson (1976). The finding of a negative relation between the activity of the strap muscles and the cricothyroid has been reported previously in other EMG studies (Atkinson, 1973; Collier, 1975; Erickson, 1976). The positive relationship between the activity of the CT and the GH has not been explored previously. We are currently investigating this with respect to possible physiological correlates of stress in English.

Relationship of muscle activity to F_0 levels. In order to ascertain the relationship of these muscles to F_0 levels, we concentrated on two key syllables (the intonational "head" and "nucleus") in two repetitions of the sentence "It's nothing less than a masterpiece," spoken on five intonational patterns [see Liberman (1975)]. We compared root mean square of the integrated EMG activity for these syllables to mean F_0 100 msec later. This delay appears to be approximately appropriate for the contraction time of the laryngeal muscles (Sawashima, 1974; Atkinson, in press). The RMS values were calculated from EMG recordings sampled at every 5 msec. The F_0 values were calculated from the voiced part of the syllable.

The GH activity has a clear positive relationship to F_0 above about 105 Hz. This interesting observation leads to speculation about possible GH function as an auxiliary pitch-raising mechanism for high F_0, when the CT needs an extra "boost" to raise F_0, as in stressed syllables. We are investigating this further. The CT activity shows a positive relation to F_0, and in this respect agrees with several other EMG studies, (for example, Atkinson, 1973, Collier, 1975, Erickson, 1976). The results are shown in Figure 3. The SH activity shows a tendency towards a negative correlation with F_0, and this too agrees with the findings of other studies (Atkinson, 1973, Collier, 1975, Erickson, 1976).

FIGURE 2

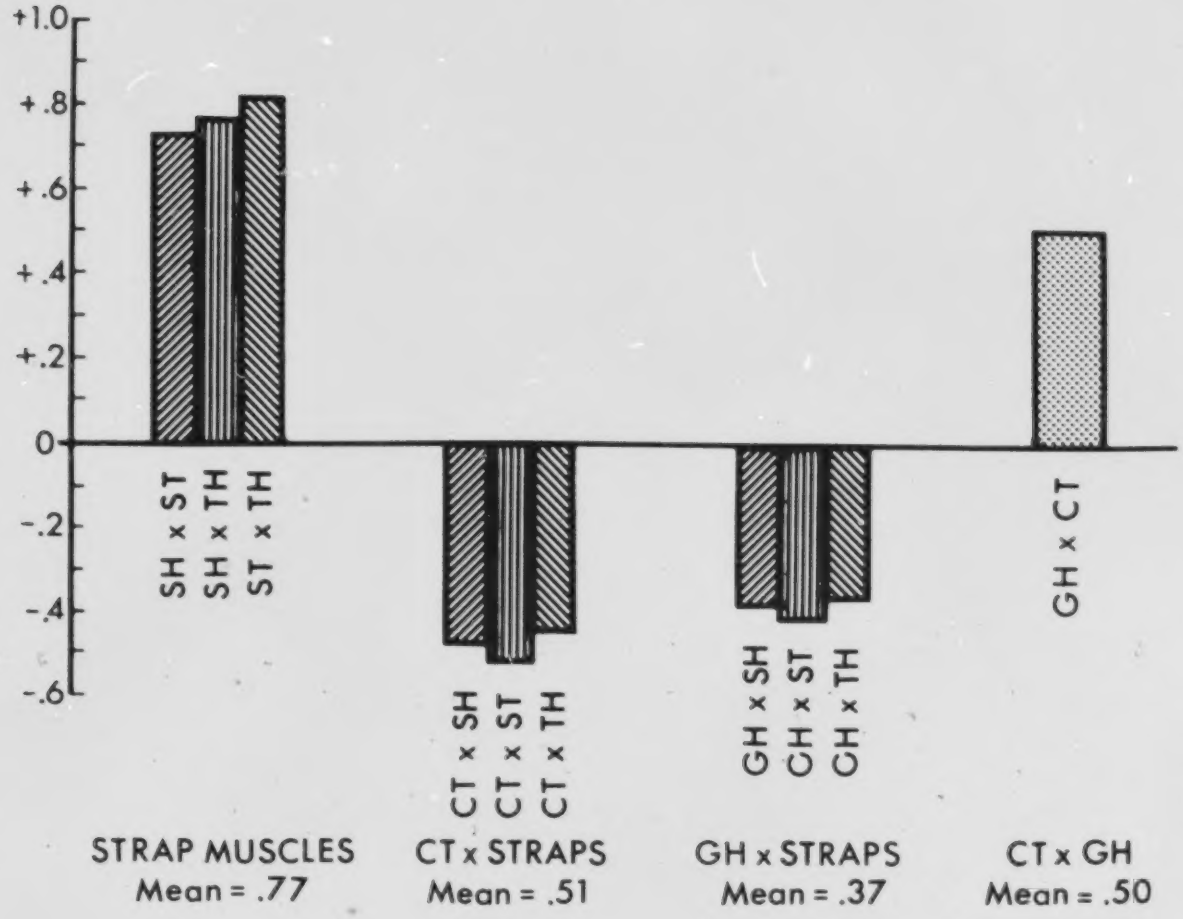


Figure 2: Correlation coefficients for the activity of the various muscles as spoken by one speaker on two repetitions of eight sentences with varying intonational patterns.

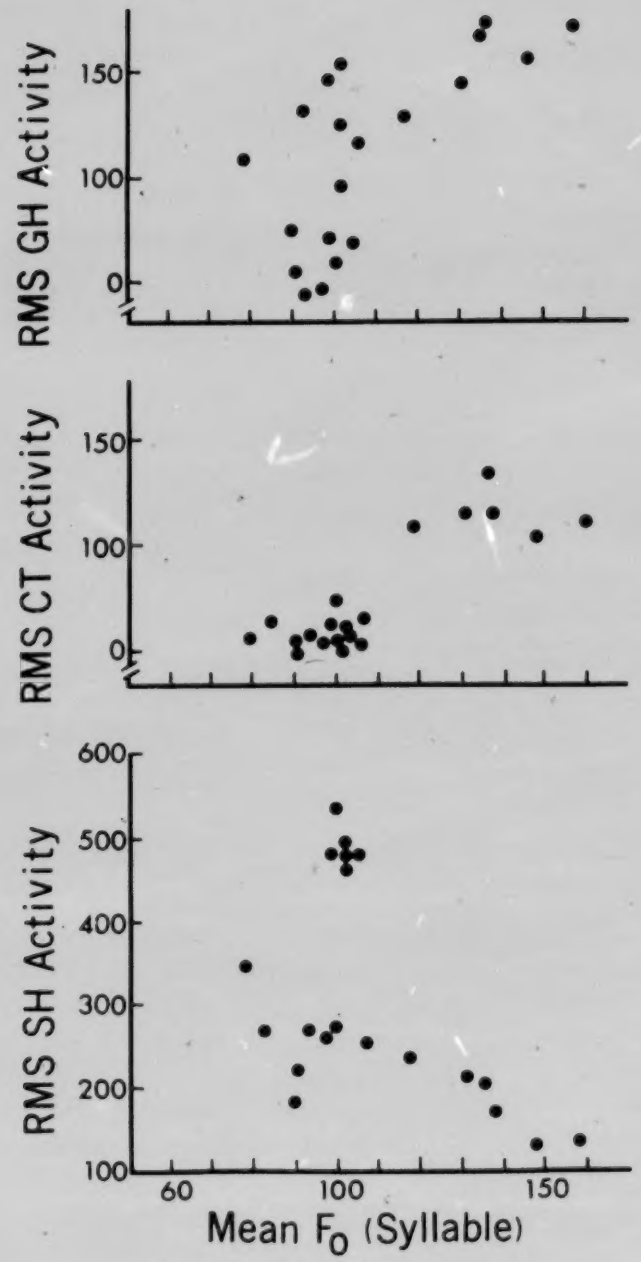


Figure 3: Relation of muscle activity in terms of root mean square values to mean F_0 of syllable. Measurements were made on the two key syllables (the intonational "head" and "nucleus") in two repetitions of the sentence "It's nothing less than a masterpiece," spoken by a single speaker on five intonational patterns. 107

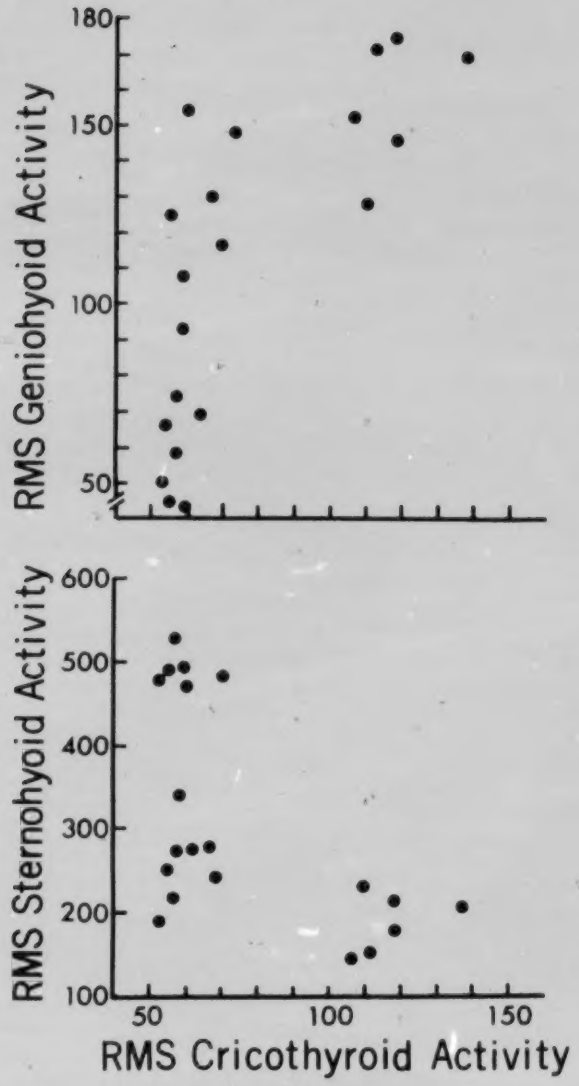


Figure 4: Relation of muscle activity in terms of root mean square values for the same syllables examined in Figure 3.

In addition to looking at the relationship of EMG activity to F_0 level, it is also interesting to look at the EMG activity for the particular syllables as shown in Figure 4. The relationship between GH activity and CT activity has not been explored previously, but holds considerable interest for future investigation. The relationship between CT activity and SH activity appears similar to that shown for Thai syllables (Erickson, 1976).

Physiological Implications

These findings can be related to a general picture of the motion of the larynx during changes in F_0. There appears to be a tendency for the larynx as a whole to move in an arc, as shown by the arrow in Figure 1: motion forward and up being generally associated with pitch raising, and motion back and down with pitch lowering. This seems to be substantiated by cineradiographic evidence that shows the hyoid bone moving up and forward during high pitch (Faaborg-Anderson and Sonninen, 1960; Colton and Shearer, 1971). (This, of course, is dependent on the head position and holds true only when the head is in the upright position seen in Figure 1). The result of this upward and forward motion of the hyoid bone is to pull the thyroid cartilage up and forward. Since the cricoid cartilage would tend to remain relatively fixed (due to its connection with the constrictor muscles and the trachea), it would tend to resist the forward component of the motion. The result of this relative motion (rotation or translation) between the two cartilages at the cricothyroid joint would tend to lengthen the vocal folds, as the larynx moves up and forward, and relax them as it moves back and down. A paper describing this view in more detail is now in preparation.

As a final remark, we wish to say that this is a preliminary investigation and the speculations and physiological findings introduced in this paper are being explored further with a view toward application to current theories of intonation and stress in English.

REFERENCES

Atkinson, J. E. (1973) Aspects of intonation in speech. Implications from an experimental study of fundamental frequency. (Unpublished Ph.D. Dissertation, University of Connecticut).

Collier, R. (1975) Physiological correlates of intonation patterns. J. Acoust. Soc. Am. 58, 249-255.

Colton, R. H. and W. Shearer. (1971) Hyoid position as a function of fundamental frequency in the modal and falsetto registers. Department of Otorhinolaryngology Laboratories and Clinics, (Experimental Phonetics Laboratory, State University of New York, Upstate Medical Center, Syracuse New York) Technical Report No. 9, November.

Erickson, D. M. (1976) A physiological analysis of the tones of Thai. (Unpublished Ph.D. dissertation, University of Connecticut).

Faaborg-Andersen, K. and A. Sonninen. (1960) The function of the extrinsic laryngeal muscles at different pitches. Acta Otolaryngol. 51, 89.

Liberman, M. L. (1975) The intonational system of English. (Unpublished Ph.D. Dissertation, M.I.T.).

Ohala, J. J. (1970) Aspects of the control and production of speech. Working Papers in Phonetics. (Univ. of California at Los Angeles) 15.

Ohala, J. J. and H. Hirose. (1970) The function of the sternohyoid muscle

in speech. (Research Institute of Logopedics and Phoniatrics, University of Tokyo Annual Bulletin) 4, 41-44.
Sawashima, J. (1974) Laryngeal research in experimental phonetics. In Current Trends in Linguistics, 12, ed. by T. A. Sebeok et al. (The Hague: Mouton), pp. 2303-2348.

Syllable Synthesis*

Ignatius G. Mattingly†

ABSTRACT

A scheme for synthesis by rule based on the phonetic syllable is described. A syllable-feature specification of the utterance to be synthesized determines a pattern of articulatory influences; these influences in turn determine the parameter values of the synthesizer.

For quite a long time, as the first slide (Figure 1) may remind you, my colleagues at Haskins Laboratories have been insisting that speech is a code, and that the encoding unit is the phonetic syllable. As the result of the merging of various coarticulatory influences, the correlates of the phonemes at the acoustic level are, in the vivid phrase of Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967), "overlapped or shingled, one onto another," yielding "irreducible segments of approximately syllabic dimensions." This observation should, indeed, be generalized to include the articulatory level as well (MacNeilage and DeClerk, 1968). In this view of the syllable [which of course goes back at least to Stetson (1951)], my colleagues have been encouraged by the findings of Koshevnikov and Chistovich (1965).

But the appeal of the phonetic syllable as an encoding unit does not rest merely on empirical observations as to the unsegmentability of anything smaller. There is not time to make the theoretical case for the syllable at length, but I would at least point out how nice it would be if it were possible to order freely the units of an ideal phonetic transcription at each prosodic level. Because of phonotactic restrictions, this condition is clearly out of the question if these units are conventional phonetic segments, but seems quite reasonable if the units are phonetic syllables. Though overlap between the physical manifestations of adjacent syllables occurs, the principle of free ordering in the transcription will be preserved as long as such overlap is predictable from the specification of the individual syllables.

From this point of view, the syllable is a cyclic process, passing from onset to peak to offset as the vocal tract moves from a more closed to a more open to a more closed configuration. The process can be realized in many

---

*This paper was presented at the 92nd Meeting of the Acoustical Society of America, San Diego, California, 15-19 November 1976.

†Also University of Connecticut.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 1977]

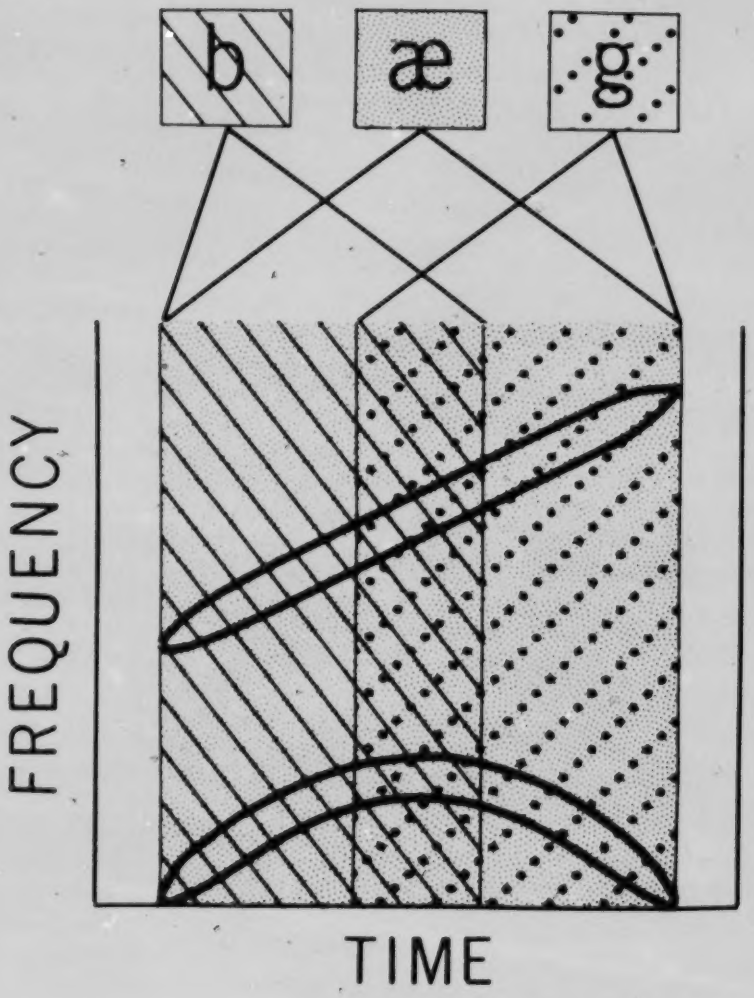


Figure 1: Interaction of consonantal and vocalic influences in [bæg]. After Liberman (1972).

different ways, depending upon the phonetic choices of the speaker. In the ideal phonetic transcription of an utterance, these choices are the values assigned to syllable features of the sort suggested by Fujimura[1] (1975, 1976), and phonetic segments have no formal status. Thus, not only the difference between [pa] and [ba], but also the difference between [pa] and [pla] depend upon a feature selection. The articulatory and acoustic consequences of a particular choice are determined by the syllabic process and may, in principle, extend throughout the physical manifestation of the syllable.

I hope it is clear that I am speaking of phonetic rather than of phonological syllables. If there are such things as phonological syllables--the matter is unsettled--they do not in general correspond one-to-one with phonetic syllables, any more than phonemes correspond one-to-one with phones. And, by the same token, if phonological syllables do not exist, the case for phonetic syllables is unaffected. A representation in terms of syllable features at the phonetic level is *a priori* entirely consistent with a segmental representation at the phonological level, and would not necessarily entail any fundamental revision of generative phonology. One of the motivations for generative phonology, in fact, is that phonological units do not necessarily correspond with phonetic units (Chomsky, 1964).

What I have been saying places a heavy explanatory burden on the concept of the phonetic syllable, and it will be credible only to the extent that syllabic processes can be shown to be orderly and more explicit. Thus, the case for the syllable will be enhanced if phonotactic restrictions, as well as much of what is now regarded as allophonic variation, can be interpreted as arising naturally from inherent properties of syllables. Synthesis by rule is an attractive tool for this undertaking.

Recently, we have begun work at Haskins on a new synthesis-by-rule program. In this new scheme, the phonetic syllable has a central role. The input to the synthesis program is a phonetic transcription of an utterance in syllabic, rather than segmental feature values. At present, the features are binary, which simplifies the transcription, but we have no strong commitment to binarity at the phonetic level. A series of ordered rules relates the feature values of the transcription to the variables used in the routine that calculates parameter values for Rod McGuire's software simulation of OVE III (Liljencrants, 1968). Since the program has as yet no phonology, it is not at present a practical vehicle for synthesizing quantities of text. Nor has consideration yet been given to stress and intonation, though the syllable plays a crucial role in these matters.

In the routine for calculating parameter values, the character of a syllable is considered to be determined by numerous influences: the vowels of the previous, current and following syllables, the final consonants of the previous and current syllables, and the initial consonants of the current and following syllables. With each such influence is associated a set of target

---

[1] O. Fujimura (1976) Syllable as the unit of speech synthesis. Unpublished memorandum, Bell Laboratories.

parameter values and a curve that represents the extent of the influence over time. An influence curve is a modified exponential function of the form $\kappa e^{\beta\tau}$. In this function [similar to the one used by Lindblom (1963) in his well-known model of consonant-vowel coarticulation], the coefficient κ determines the effective time of onset of an influence, and the exponent determines its rate of growth. On phonetic grounds, these properties of the function are appealing, since one would expect both the shape and the relative timing of the various influence curves to be significant variables. Of course, there are other functions that might possibly have been used instead. The value of the function is restricted to the range 0...1, since it is used as a weight, and at a certain time $\tau(x)$ after the notional beginning of the syllable cycle, β becomes negative, so that the influence will begin to diminish. The target values, and the values of κ, β, $\tau(x)$, and other variables are assigned by the rules.

It might be objected that the notion of an "influence" simply reintroduces the phonetic segment in a new guise, particularly when I refer to the influences of consonants and vowels, and employ the conventional terms for manner classes. But unlike phonetic segments, influences are not linearly ordered; their temporal relationship is more complex than that. And "consonant," "vowel," and the various manner class terms are to be understood not as segment categories, but as labels given to various recognizable aspects of the syllabic cycle by which they are defined.

Because of our particular interest in the temporal patterns of events within the syllable, we have provided various ways to control these patterns in the program.[2] As we have just seen, κ controls the effective onset of an influence; by manipulating this variable, different degrees of consonantal and vocalic coarticulation may be provided. Since the moment when an influence begins to diminish is a variable, articulatory holds for stops and fricatives can be represented. Moreover, each influence can potentially increase the duration of the syllable by a certain amount. If such an increment is called for, the onsets of syllable-final and following-syllable influences are postponed by appropriately reducing their κ values.

The actual parameter values for a particular 5-msec sample of speech are derived by an iterative calculation. The influences are regarded as ordered, from vowels to fricatives to stops. At each iteration, the value computed for a parameter is

$$V_i = V_{i-1} + I_i(T_i - V_{i-1})$$

that is, the weighted sum of the target values associated with the influence and the value computed at the preceding iteration, the relative weighting being determined by the value of the influence function at that point in the syllable. (At the first iteration, the target value for the vowel of the previous syllable serves as the seed value V_0.) Because of the large number of influences, the burden of calculation would be considerable were it not

---

[2] Our investigations of syllable duration are reported in a paper read by Linda Shockey at an earlier session (Shockey and Mattingly, 1976).

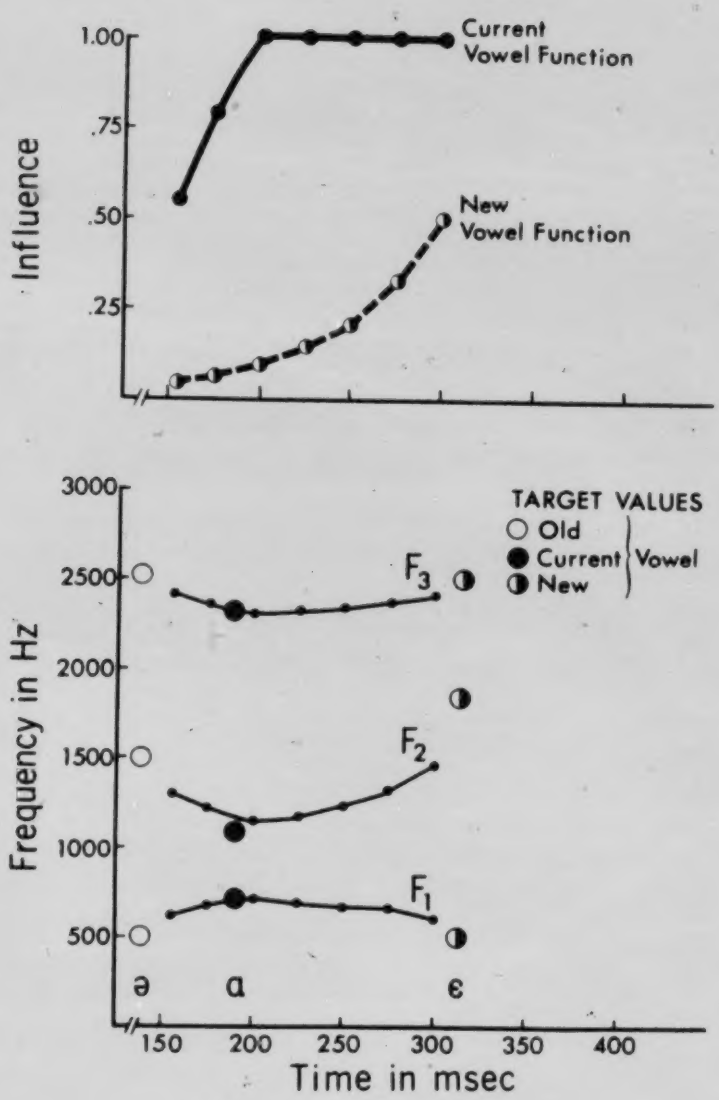


Figure 2: Influence functions for current vowel and new vowel for [ɑ] preceded by [ə] and followed by [ɛ], and resulting formant trajectories.

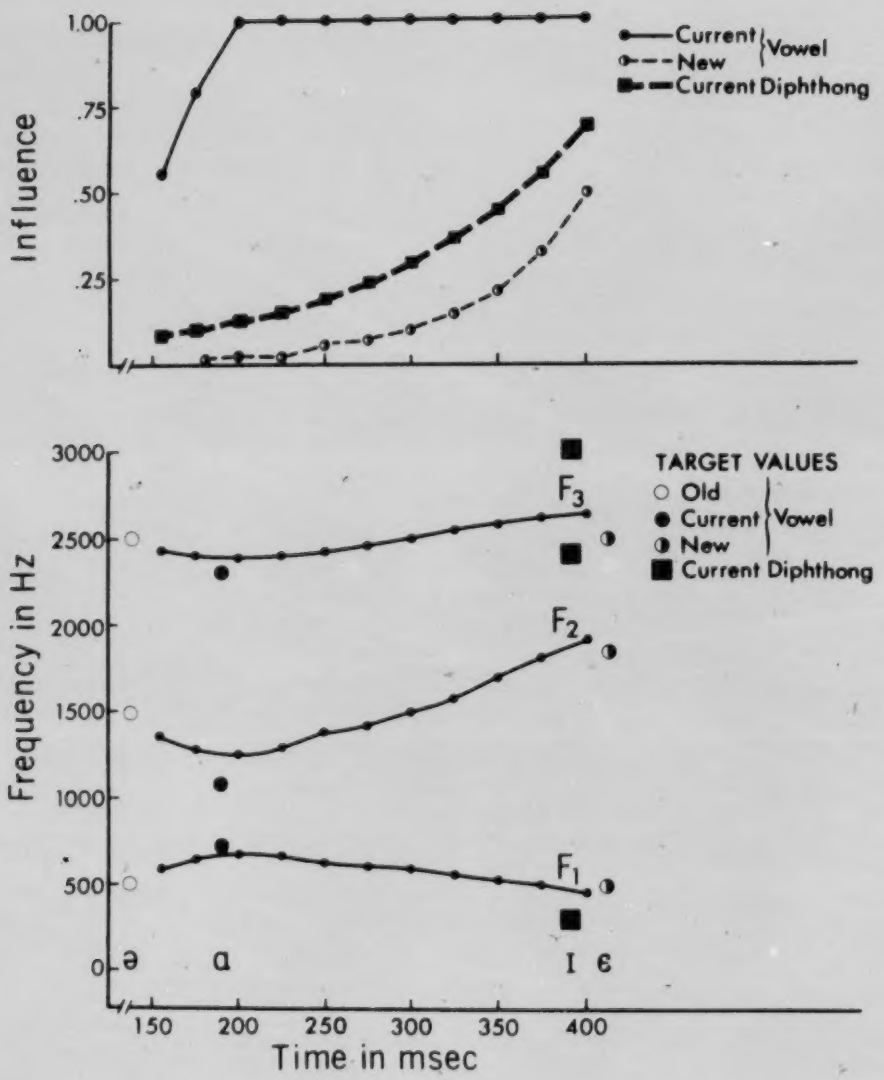


Figure 3: Effect of adding influence of diphthong on syllable shown in Figure 2.

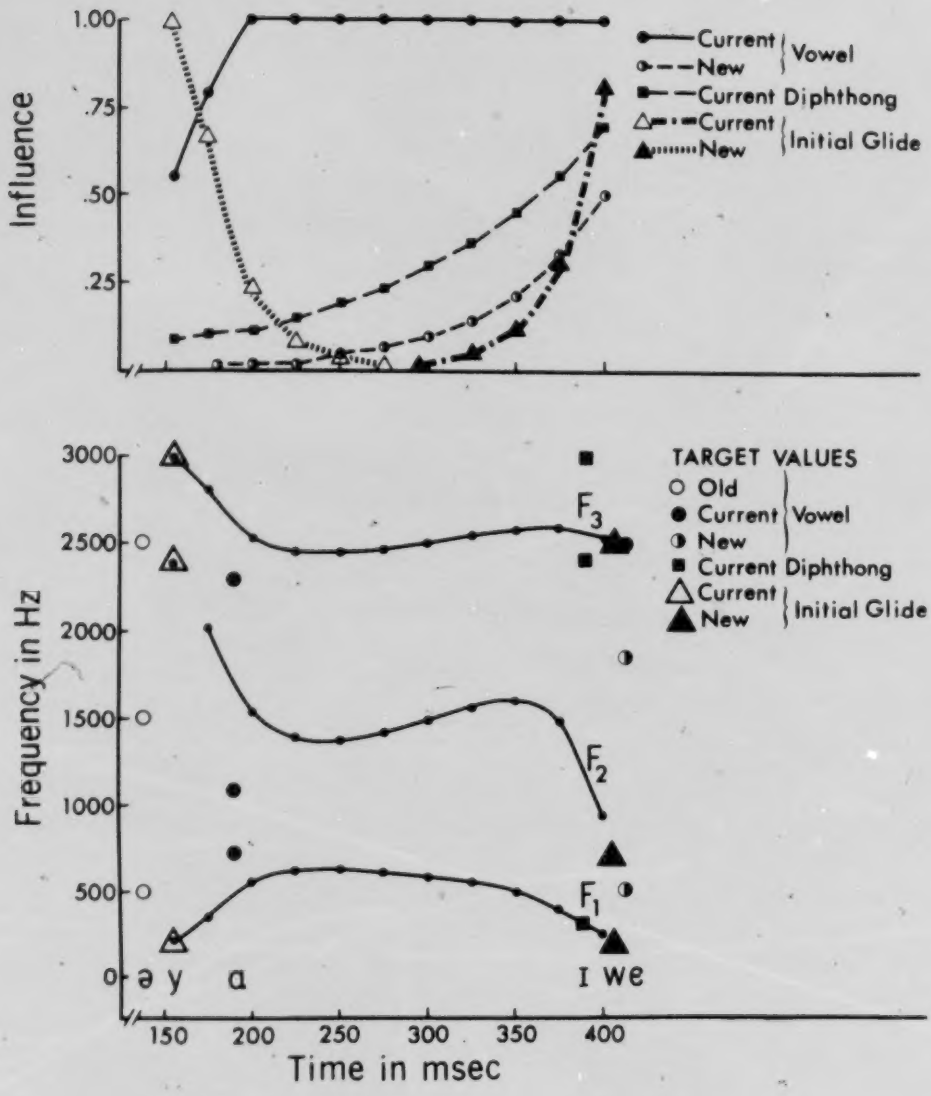


Figure 4: Effect of adding influence of current initial glide and following initial glide on syllable shown in Figure 3.

that at any one time, many potential influences are inactive, that is, have values near zero, and may simply be ignored.

The next three slides (Figures 2, 3, 4) illustrate how the overlapping of various influences is realized. This slide (Figure 1) shows an [ɑ] assumed to have been preceded by [ə] and followed by [ɛ]. The curve with black circles in the upper portions of the slide shows the increasing influence of the [ɑ] at the expense of the [ə] of the preceding syllable. The curve with white circles shows the increasing influence of the [ɛ] of the following syllable at the expense of the [ɑ]. The lower portion of the slide (Figure 1) shows the target formant frequencies for all three vowels and the formant movements resulting from their influence.

In this slide (Figure 3) the influence of a final palatal glide is interposed, in addition to the other influences, to give the diphthong [ɑɪ], and the formants change accordingly.

Finally, in Figure 4, the influences of an initial [y] glide in the [ɑɪ] syllable and of an initial [w] glide in the following syllable are superimposed upon the other influences.

This way of calculating parameter values will be recognized as a generalization of the method used by Holmes, Mattingly and Shearme (1964) and by the earlier Haskins programs for calculating formant transitions (Mattingly, 1968a, 1968b; Kuhn, 1973), in which the "boundary value" used as a basis for interpolation was the weighted sum of the target frequencies of two adjacent phones. It is also analogous, as Tim Rand has pointed out, to a series of filters, each of which corresponds to an "influence."

The scheme, as described so far, is quite general, and could be implemented in terms of articulatory gestures, or vocal tract shapes, or formant movements, depending upon the choice of parameters. The most interesting and satisfying implementation would be the articulatory one, but because we are anxious to explore temporal questions as soon as possible, we are beginning with an acoustic version.

REFERENCES

Chomsky, N. (1964) Current Issues in Linguistic Theory. (The Hague: Mouton).

Fujimura, O. (1975) Syllable as a unit of speech recognition. IEEE Trans. Acoustics Speech and Signal Processing ASSP-23, 82-87.

Fujimura, O. (1976) Syllables as concatenated demisyllables and affixes. J. Acoust. Soc. Am. 59, S55(A).

Holmes, J. N., I. G. Mattingly, and J. N. Shearme. (1964) Speech synthesis by rule. Lang. Speech 7, 127-143.

Kozhevnikov, V. A. and L. A. Chistovich. (1965) Speech Articulation and Perception, trans. from Rech' Artikulyatsiya, i vospriyatiye (Moscow-Leningrad). (Washington, D.C.: Joint Publications Research Service).

Kuhn, G. M. (1973) A two-pass procedure for synthesis by rule. J. Acoust. Soc. Am. 54, 339(A).

Liberman, A. M. (1972) The specialization of the language hemisphere. In The Neurosciences Third Study Program, ed. by F. O. Schmitt and

F. G. Worden. (Cambridge, Mass.: MIT Press).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psych. Rev. 74, 431-461.

Liljencrants, J. C. W. A. (1968) The QVE III speech synthesizer. IEEE Trans. Audio and Electroacoustics AU-16, 137-140.

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1771-1781.

MacNeilage, P. F. and J. L. DeClerk. (1968) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Am. 45, 1217-1233.

Mattingly, I. G. (1968a) Synthesis by Rule of General American English, Supplement to Haskins Laboratories Status Report on Speech Research, April.

Mattingly, I. G. (1968b) Experimental methods of synthesis by rule. IEEE Trans. Audio Electroacoust. AU-16, 198-202.

Shockey, L. and I. G. Mattingly. (1976) Temporal Patterns of Syllables. J. Acoust. Soc. Am. 60, S26-27.

Stetson, R. H. (1951) Motor Phonetics. 2nd ed., (Amsterdam: North Holland).

Articulatory Movements in VCV Sequences

Thomas Gay†

ABSTRACT

The purpose of this experiment was to study both the timing and positional properties of articulatory movements in VCV utterances. Conventional cinefluorographic techniques were used to track the movements of the upper lip, lower lip, jaw, tongue tip, and tongue body of two speakers who read randomized lists of VCV utterances containing the vowels /i,a,u/ and the consonants /p,t,k/, in all possible combinations. Results showed that the timing of articulatory movements in a VCV sequence is constrained by the intervocalic consonant, even if the gesture for the consonant is not a contradictory one. Anticipatory movements toward the second vowel always begin during the closure period of the intervocalic consonant. The appearance of carryover coarticulation effects depends on the phonetic identity of the particular segment or degree of involvement of the articulator. Carryover effects, like anticipatory effects, did not extend beyond an immediately adjacent segment. These findings suggest that the rules governing the segmental input to a speech string might be simpler than present models suggest.

INTRODUCTION

The purpose of this paper is to explore a number of questions related to the properties of articulatory movements in VCV utterances. The experiment was motivated by the fact that in the literature there exist contradictory reports concerning the nature and extent of various coarticulatory phenomena. While the traditional view, and the earlier papers of Öhman (1966), and Daniloff and Moll (1968), for example, hold that coarticulation is inherent in the programming of speech sequences, and that its effects can extend across various structural boundaries, other more recent studies (Gay, 1974a, Gay, 1974b; Bell-Berti and Harris, 1975) suggest that the rules governing

---

†Also University of Connecticut Health Center, Farmington, Conn.

Acknowledgment: The author wishes to thank Dr. J. Daniel Subtelny, Department of Orthodontics, Eastman Dental Center, Rochester, New York, for use of the cinefluorographic facilities at Eastman, and Ms. Kathleen Kirchmeier for her assistance in the analysis of the data. This research was supported by grants from the National Institute of Neurological and Communicative Disorders and Stroke (NS-10424), and The National Science Foundation (GSOC-7403725).

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

coarticulation (both anticipatory and carryover) might be somewhat simpler than previously believed.

Anticipatory coarticulation effects are essentially timing effects: movements toward some parts of a feature target of a given segment begin before others. In a study of anticipatory lip rounding, Kozhevnikov and Chistovich (1965) found that the onset of the rounding gesture for the vowel /u/ placed in a CCV syllable occurred at the beginning of the syllable. Daniloff and Moll (1968), in extending the observations of Kozhevnikov and Chistovich, showed that lip rounding for /u/ can begin across as many as four segments ahead of the vowel. In their experiment, anticipation of lip rounding for the vowel /u/ was studied for a number of mono- and disyllabic single and two-word utterances embedded in sentence frames using lateral view cinefluorography. Onset of lip rounding usually began during the closure phase of the first consonant in the sequence, and was not affected by the position of word or syllable boundaries within the sequence. Another type of anticipatory coarticulation was shown to exist by Öhman. In a spectrographic study of coarticulation in VCV sequences, Öhman showed that the variability observed in transition movements to the consonant could be predicted by the formant frequencies of the second vowel. This led Öhman to conclude that vowel-to-vowel movement in a VCV is essentially diphthongal with the consonant simply superimposed on the basic gesture; in other words, movements toward the second vowel begin independently from and at about the same time as those toward the consonant. In other studies, Moll and Daniloff (1971) showed that velopharyngeal opening for a nasal consonant can begin two vowels in advance of the consonant, and McClean (1973) showed that in a CVVN sequence, velar opening for the final nasal begins ahead of the syllable boundary, unless the two vowels are separated by a marked junctural boundary. These studies, among others, suggest that articulatory encoding is a complex phenomenon whose effects can spread across several adjacent segments. Most studies support, either explicitly or implicitly, Henke's (1966) articulatory model that proposes the operation of a mechanism that scans future segmental inputs, or features thereof, and sends commands for the immediate attainment of those feature targets that would not interfere with the attainment of immediately intervening articulations.

However, in several recent studies, both electromyographic (Gay, 1974b; Ushijima and Hirose, 1975) and acoustic (Ohde and Sharf, 1974; Bell-Berti and Harris, 1975), evidence was used to argue against the ubiquity of anticipatory coarticulation effects in speech. In an experiment by Gay (1974b), EMG recordings were obtained from the genioglossus and orbicularis oris muscles of two subjects during the production of various VCV syllables. In those utterances where the genioglossus muscle was involved in the production of both the first and second vowels (as in /upi/ or /itu/), or where the first and second vowels were the same (as in /ipi/ or /utu/), a cessation of activity occurred for the genioglossus muscle during the time of consonant production. In other words, each vowel in the sequence (even in a symmetrical VCV) was marked by a separate muscle pulse. The interpretation of the finding reflected a discontinuity in vowel-to-vowel movement, and thus, a contradiction to Öhman's (1966) diphthongal movement hypothesis. Another finding of this experiment was the presence of a trough in the orbicularis oris envelope during the production of an alveolar or velar consonant that separated two rounded vowels. This finding was not consistent with others

that showed a considerably earlier onset of the lip rounding gesture (Kozhevnikov and Chistovich, 1965; Daniloff and Moll, 1968; Benguerel and Cowan, 1974). In another EMG experiment, Ushijima and Hirose (1975) showed that in a CVVN sequence, lowering of the velum in anticipation of the final nasal was restricted by the syllable boundary. While these results were obtained from Japanese, they nonetheless argue against a general model of anticipatory velar lowering.

In an experiment performed by Bell-Berti and Harris (1975), spectrographic measurements were made from eighteen utterance types that consisted of the vowels /i,a,u/ in CVC combinations with the consonants /p,t,k/. The data showed that the effects of the terminal consonant on the midpoint of the stressed vowel were not nearly as large as those of the initial consonant; in other words, the carryover effect of the initial consonant on the vowel is considerably greater than the anticipatory effect of the second consonant. The same results were also obtained independently by Ohde and Sharf (1974): in a variety of CVC sequences, carryover articulation effects on vowel targets were likewise greater than anticipatory effects.

Carryover coarticulation effects are essentially positional effects and exist in the form of variability in target (or target feature) positions as a function of changes in phonetic context. Carryover effects have traditionally been attributed to mechanical or inertial effects and, in general, have been studied less extensively than anticipatory effects. Although carryover effects have been shown to exist at both the EMG and articulatory levels (MacNeilage and DeClerk, 1969; Sussman, MacNeilage, and Hanson, 1973; Gay, 1974c), the pervasiveness of these effects is somewhat in doubt. In a study of the production of thirty-six CVC monosyllables, MacNeilage and DeClerk (1969) found that some aspect of the production of every phone was always influenced by a preceding phone and almost always influenced by a following phone. In particular, the size of the EMG signal would be different depending on the identity of the adjacent vowel or consonant. In countering the argument that a motor command representation of the phone shows less variability than an articulatory target representation, MacNeilage (1970) later proposed that the observed EMG variability reflected a complex motor strategy, the underlying goal of which is a relatively invariant articulatory end. The concept of an articulatory-based target system as proposed by MacNeilage was further supported, at least for vowels, by the cinefluorographic data of Gay, Ushijima, Hirose, and Cooper (1974) and Gay (1974a). In the latter study, lateral view x-ray motion pictures were obtained from two speakers who produced the vowels /i,a,u/ in a variety of VCV contexts. The results of this experiment showed that for both subjects, the target positions for both /i/ and /u/, in both pre- and postconsonantal positions, remained quite stable (within 2-3 mm) across changes in the consonant and transconsonantal vowel. Finally, a careful examination of Öhman's (1966) acoustic data shows that carryover effects of the first vowel or the intervocalic consonant on the formant frequencies of the second vowel were virtually nonexistent: formant frequencies fell within a 50-60 Hz range regardless of the identity of the preceding phones. However, in contrast to the studies cited above, carryover effects have been shown to exist at the articulatory level. Sussman, MacNeilage, and Hanson (1973) and Gay (1974c), for example, have produced data showing jaw position during consonant and vowel production to be sensitive to the degree of jaw opening of an adjacent

phone. Thus, although evidence exists to support an articulatory target formulation, no present theory specifies the rules governing failure to achieve a particular target.

The divergent research results of the last ten years, whether arising from differences in interpretation or the utilization of different experimental techniques, nevertheless serve to point out that a number of important questions concerning the dynamic properties of speech gestures remain unanswered. In this experiment, both the timing and positional properties of articulatory movements in VCV utterances were studied, using conventional pellet tracking and spectrographic techniques, in an attempt to provide answers to some of these questions. The format of the experiment was designed to explore questions related to two particular issues: 1) the constraints an intervening consonant might place on the movements of the articulators, especially the tongue body, from one vowel to another (is the movement from vowel to vowel essentially diphthongal or is it locked somehow to the intervocalic consonant?) and 2) the extent of carryover coarticulation effects throughout the syllables (are such effects limited to phonetically unmarked features such as jaw position or do they extend to other properties of both vowel and consonant production?).

METHOD

Subjects and Speech Material

Subjects were two adult males, both native speakers of American English. The speech material consisted of CVCVC strings where the initial and final consonants remained constant (/k/ and /p/, respectively), and the medial VCV sequences contained the vowels /i,a,u/, and the consonants /p,t,k/ in all possible combinations. Each of the twenty-seven utterances was placed in the carrier phrase, "Say _____ again," and random-ordered into a master list.

Data Recording

Lateral view x-ray films were recorded with a 16 mm cine camera at a speed of 60 fps. The x-ray generator delivered 1 msec pulses at 120 kv to a nine-inch image intensifier tube. For purposes of tracking articulatory movements, 2.5 mm lead pellets were attached to the upper and lower lips, tongue tip, dorsum, and body (at two locations) of both subjects.[1] In addition, a reference pellet was attached at the embrasure of the upper central incisors. Jaw movements for both subjects were tracked by measuring the distance between the tip of the lower central incisors and the reference pellet. All pellets were attached at the midline using a cyanoacrylate adhesive. The locations of the pellets are shown for both subjects in Figure 1.

Each subject was positioned in a head holder. The subjects were instructed to read the list at a comfortable speaking rate and with equal

---

[1]The second, more posterior, tongue body pellet for Subject GNS fell off during the experiment run.

FIGURE 1

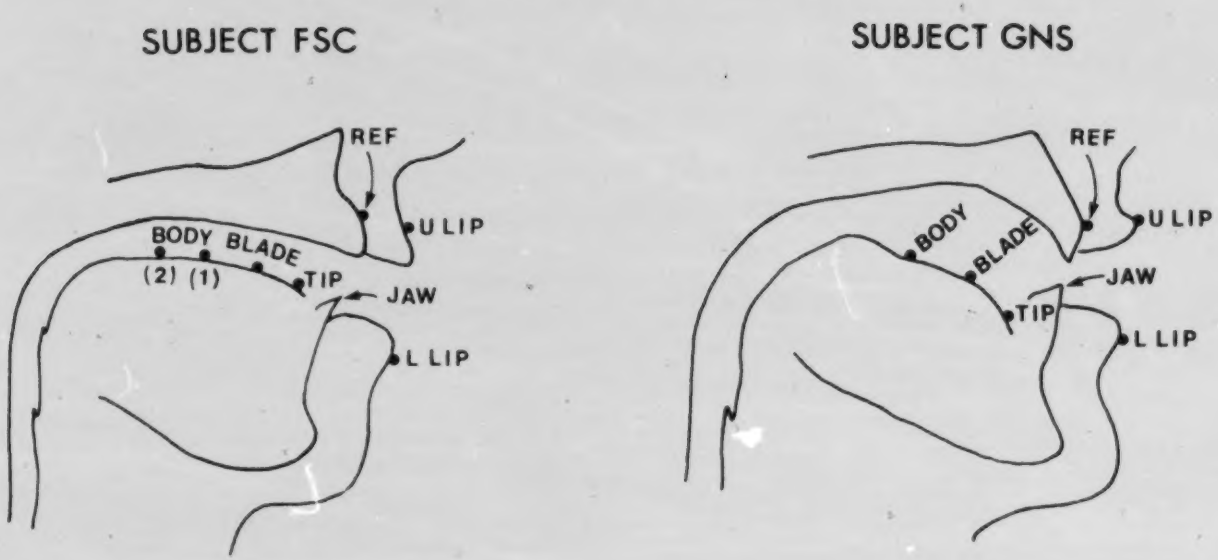


Figure 1: Locations of pellets for tracking articulatory movements. Jaw movements were measured at tip of lower central incisors.

stress placed on the two syllables. A brief practice session preceded each run. During the x-ray run, the corresponding acoustic signal was also recorded on magnetic tape.

Data Analysis

A semiautomated system for analyzing the x-ray data was developed for this purpose. It consists essentially of a 16 mm film analyzer (Perceptoscope, Mark III) and digitizing tablet (Summagraphics) that is interfaced to a small laboratory computer (D.E.C., PDP/8E). The film image is projected, frame-by-frame, via an overhead mirror system onto the surface of the digitizing tablet. The position coordinates of each pellet (or other anatomical landmark) are stored in the computer when a hand-held pen is depressed over the pellet location. Sections of the tablet outside the image area are used for control operations, for example, storing a special skip code or indicating end of utterance. The computer measures the X and Y coordinate positions of each pellet relative to the position of the reference pellet and stores the accumulated data, frame-by-frame-by-utterance, on disk. A second program is used to display the X and Y components separately as a movement track on a large display scope. The resolution of the digitizing tablet is .25 mm. By projecting the film twice real size, measurement error is easily reduced to within $\pm$ 1 mm. This was the usual maximum real size error obtained from repetitive measurements of selected samples.

One particular problem inherent in x-ray pellet tracking techniques is the obstacle dental fillings present in marking pellet locations. Because of the density of amalgams, the pellets become lost when they enter behind such fillings. Dental restorations interefered with the tracking of the first tongue body pellet of Subject FSC and the tongue body and tongue tip pellets of Subject GNS, both to varying degrees in different utterances.

Wide band spectrograms, using a Haskins Laboratories digital spectrograph routine, were made for all utterances. A particular advantage of this routine is a software thresholding feature that can be used to reduce the background noise produced by the x-ray generator. This permitted spectrographic measurements to be made for almost all of the vowel nuclei, although the less intense parts of the signal associated with formant transitions were lost in the noise.

The acoustic recordings of both subjects were analyzed for the purpose of determining whether stress differences appeared for the first and second vowels. Perceived destressing occurred consistently for /a/ in preconsonantal position for Subject GNS. Destressing of preconsonantal /a/ was also evident in the spectrographic measures. First and second formant frequencies for /a/, pooled across consonants and vowels, were 640 Hz and 1340 Hz for the initial position, and 810 Hz and 1210 Hz, for the final position. Instances of first vowel destressing for /a/ also occurred for Subject FSC, but not consistently. These were the only stress effects that appeared for either subject.

RESULTS AND DISCUSSION

The Timing of Articulatory Movements

One of the basic questions addressed in this experiment concerns the coordination of articulatory movements throughout a VCV utterance, that is, the relative timing of the movements of the tongue body in relation to those of the lips, jaw, and tongue tip, especially during the production of the intervocalic consonant. The three different consonants appearing in the various utterances were selected on the basis of the varying degrees of involvement of the tongue during their production: complete independence as a primary articulator for /p/, only tongue tip involvement for /t/, and complete involvement of the tongue body as a primary articulator for /k/. As will be shown, however, tongue body movements are either involved in or constrained by each of the three different intervocalic consonants.

Measurements of the relative onsets of articulatory movements in the various VCV sequences are summarized in Figure 2. This figure shows the ranges of onset times of tongue body, jaw, and primary articulator movements (either the lower lip, tongue tip or tongue body for /p,t,k/ respectively), from the first vowel to the intervocalic consonant, and from the intervocalic consonant to the second vowel. Onset times are relative to the time of closure for the consonant and are plotted separately for the three consonants. These data provide an overall picture of the relative timing of articulatory movements through the VCV sequence.

For both subjects, the timing of articulatory movements from the first vowel to the consonant were far more constrained than articulatory movements from the consonant to the second vowel. For closing movements, the onset times of tongue body, jaw, and primary articulator movements fell within the same overall time window. While the window, itself, is rather wide, coordination within the window is much more constrained, with the movements of the tongue body, jaw, and primary articulator beginning within 10-15 msec of each other. The observed overall variability could not be attributed to either the duration of consonant closure or the identity of the first vowel, although there was some tendency for earlier starting times to occur for /a/, probably as a function of greater articulatory displacement. It should also be noted that in a number of instances, notably those sequences where the first vowel is /u/, closing movements of the primary articulator were not accompanied by corresponding movements of either the tongue body or jaw.[2]

In contrast to the constrained closing movements from the first vowel to the consonant, opening from the consonant to the second vowel was character-

---

[2]When the intervocalic consonant was either /p/ or /t/, tongue body participation in the consonant gesture depended on the identity of the first vowel; tongue body movements always accompanied primary articulator movements when the first vowel was /a/, sometimes showed movement when the first vowel was /i/, and never showed movement when the first vowel was /u/. For /k/, of course, the tongue body always showed movement into the consonant. In those cases where tongue body movements did not appear for the consonant, the tongue body simply maintained the target position of the first vowel.

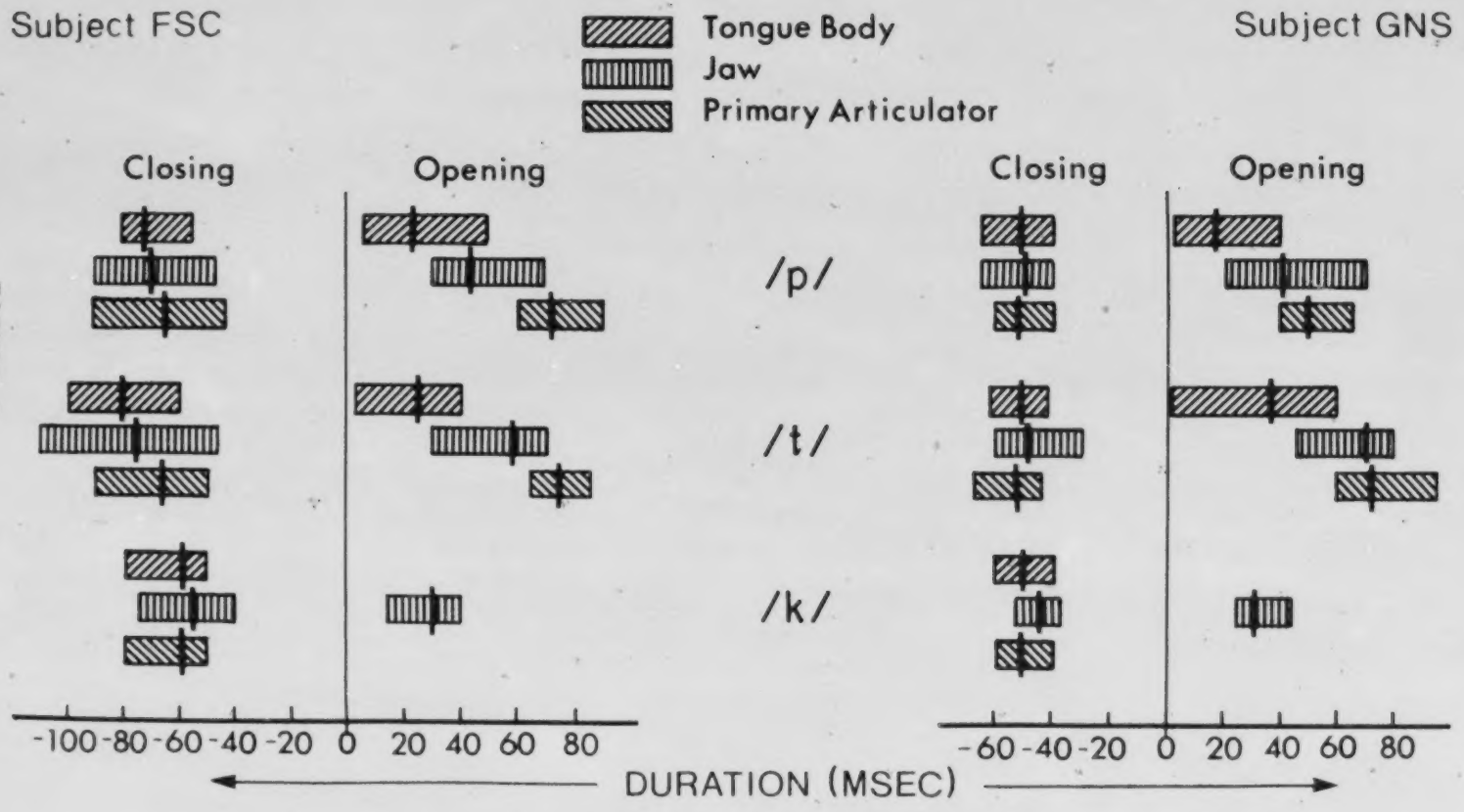


FIGURE 2

Figure 2: Ranges of relative onset times of articulatory movement associated with consonant closing and vowel opening. The vertical lines indicate mean values.

ized by a staggered pattern of movements. For both subjects, opening toward the second vowel began with the tongue body, and was followed by the jaw and primary articulator, in that order. Movements of the tongue body began anywhere from 5-50 msec for Subject FSC, and 5-60 msec for Subject GNS after the time of consonant closure. All tongue body movements, however, were underway before the time of consonant release. The onset time of jaw opening also varied within the interval of consonant closure, but usually followed tongue movements and preceded primary articulator movements. The variability of opening onset times, like those for closing, did not correspond to any feature other than a tendency for earlier opening to occur for a following open vowel.

The dynamic properties of articulatory movements in a VCV sequence, and the rules that govern those movements, will be discussed for each consonant category, using graphical illustrations produced from the frame-by-frame measurements of the x-ray films. The movements of the tongue body, lips, and jaw for a VCV sequence where the intervocalic consonant is /p/, are illustrated for both subjects in Figure 3. This figure shows the movement track of the height dimension for the sequence /ipa/. Each track was graphed from discrete points measured every film frame, that is, at approximately 17 msec intervals. Measurements begin during the closure period of the initial /k/ and end at the time of closure for the final /p/; 0 on the abscissa corresponds to the time of consonant closure. This figure illustrates the constraints that the intervocalic consonant places on the timing of the tongue body from vowel to vowel. The movement of the tongue body from the first vowel to the second vowel does not begin until after closure for the intervocalic consonant is completed. This, of course, was a salient feature in the production of all VCV utterances by both subjects (ref. Figure 2). This figure also shows that the movements of the tongue body begin ahead of those for the jaw. The delay time is approximately 40 msec for Subject FSC and 60 msec for Subject GNS. This delay suggests that tongue body movements toward the vowel are probably independent from jaw movements toward the vowel. This figure also illustrates the variability of jaw movements associated with consonant production. For Subject FSC, jaw closing begins at the time of lip closing, while jaw opening precedes lip opening. For Subject GNS, on the other hand, jaw closing does not accompany lip closing and jaw opening follows lip opening (this pattern is the only exception to the general rule). As is also evident in this figure, upper lip contributions to lip closure were negligible for both subjects. Finally, Subject FSC showed a pattern of lip closure that was often characterized by continued compression throughout the closure period.

Consonant constraints on vowel-to-vowel movements are as evident in the front-back dimension as in the height dimension. Figure 4 shows tongue movement in the X dimension plotted against the same baseline as lower lip movement in the Y dimension, both as a function of time for the sequence /ipu/. Again, it is apparent that tongue movement toward the second vowel does not begin until after consonant closure. The data for Subject GNS also show what might be a tongue body gesture associated with the consonant. Such a gesture, however, did not appear regularly in the data, nor did the tongue body appear to reach a specific, repeatable target position when such a gesture did appear.

FIGURE 3

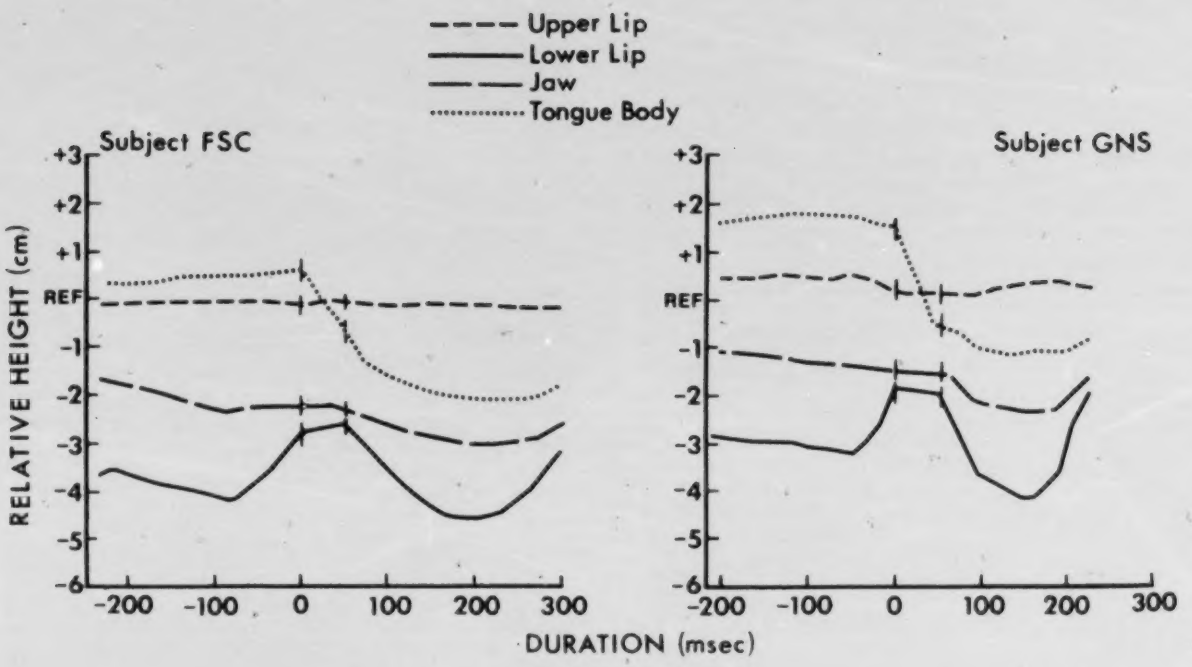


Figure 3: Movement tracks for utterance /ipa/. 0 on the abscissa, in this and all subsequent figures, corresponds to time of consonant closure; vertical bars indicate the times of consonant closure and consonant release. The tongue body pellet for Subject FSC is the second, more posterior, one.

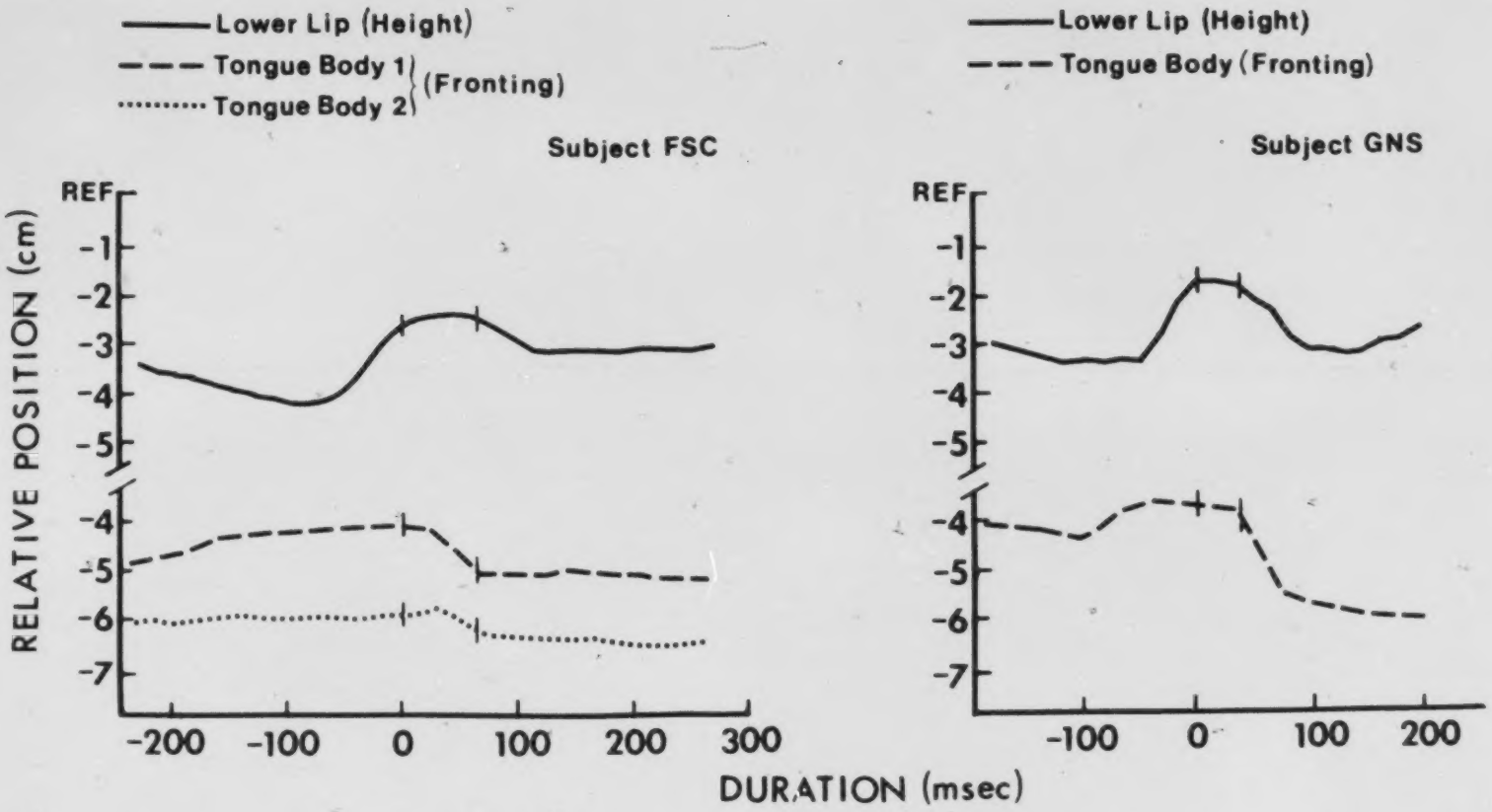


FIGURE 4

Figure 4: Movement tracks for utterance /ipu/. Both height and fronting measurements are plotted on the same baseline.

The same rules for tongue body movement associated with /p/ are also evident for utterances where /t/ is the intervocalic consonant (Figure 5). Here, as before, movements toward the second vowel do not begin until after closure for /t/. Also, this figure shows that the movements of the tongue body, tongue tip, and jaw are again, independent from each other; they all begin moving into the second vowel at different times, with the tongue body leading the jaw and tongue tip, in that order.

Perhaps the best illustration of consonantal constraints on tongue body movements is one where the first and second vowels of the utterance are the same. Figure 6 shows the movement tracks for the jaw and four tongue pellets during the production of /iti/ for Subject FSC. Instead of the tongue maintaining the /i/ target during the consonant, the tongue blade and both tongue body pellets show movement throughout the consonant gesture. The blade and anterior tongue body pellet appear to shadow movements of the tip, while the posterior tongue body pellet moves in the opposite direction (lower). Because the tongue body is displaced at least 5 mm from the vowel target during the time of consonant production, the movement is probably not passive (a pressure perturbation for example). Rather, it would seem that the gesture is a facilitory one or one that reflects a strategy to modulate the degree of aspiration that might otherwise occur if the postalveolar channel were too constricted.[3] It should also be noted that the present finding agrees with the x-ray data of Kent (1970) that also showed tongue body movement in a symmetrical VCV at the time of consonant production.

The most interesting tongue movements are those associated with /k/ production. Figure 7 shows both the height and fronting components of tongue body movement during the production of /aki/, /aka/, and /aku/, for Subject FSC. These traces show that the tongue body is in continuous movement throughout the closure phase of the consonant. From the time of /k/ closure, the tongue body continues to move upward and forward for a following /i/ or /a/, and upward and slightly backward for a following /u/. Continuous movement of the tongue body during /k/ production has also been reported in a number of other papers. The data of both Kent (1970) and Perkell (1969) show elliptical patterns of movement of the tongue body for /k/ in symmetrical /VkV/ and /əkV/ sequences, respectively. A similar pattern exists in the present symmetrical /VkV/ sequences and would emerge from the /aka/ data in Figure 7 if a composite trace were constructed from the two movement tracks. The present data are also in general agreement with those of Houde (1967) who showed that the tongue body was in continuous movement during /k/ in an asymmetrical /VkV/ sequence.

Of particular interest in the present data is the finding that, irrespective of the identity of the second vowel in the sequence, closure for /k/ occurs at approximately the same location in the vocal tract. Tongue movement continues through the consonant, with release occurring at different locations in anticipation of the following vowel (ref. Figure 7). While the three movement tracks are within 3 mm of each other, in both dimensions, at closure, they diverge towards release, at which point the differences are 8 mm between /i/ and /a/ in the height dimension, and 10 mm between /i/ and /u/

---

[3] K. N. Stevens: personal communication.

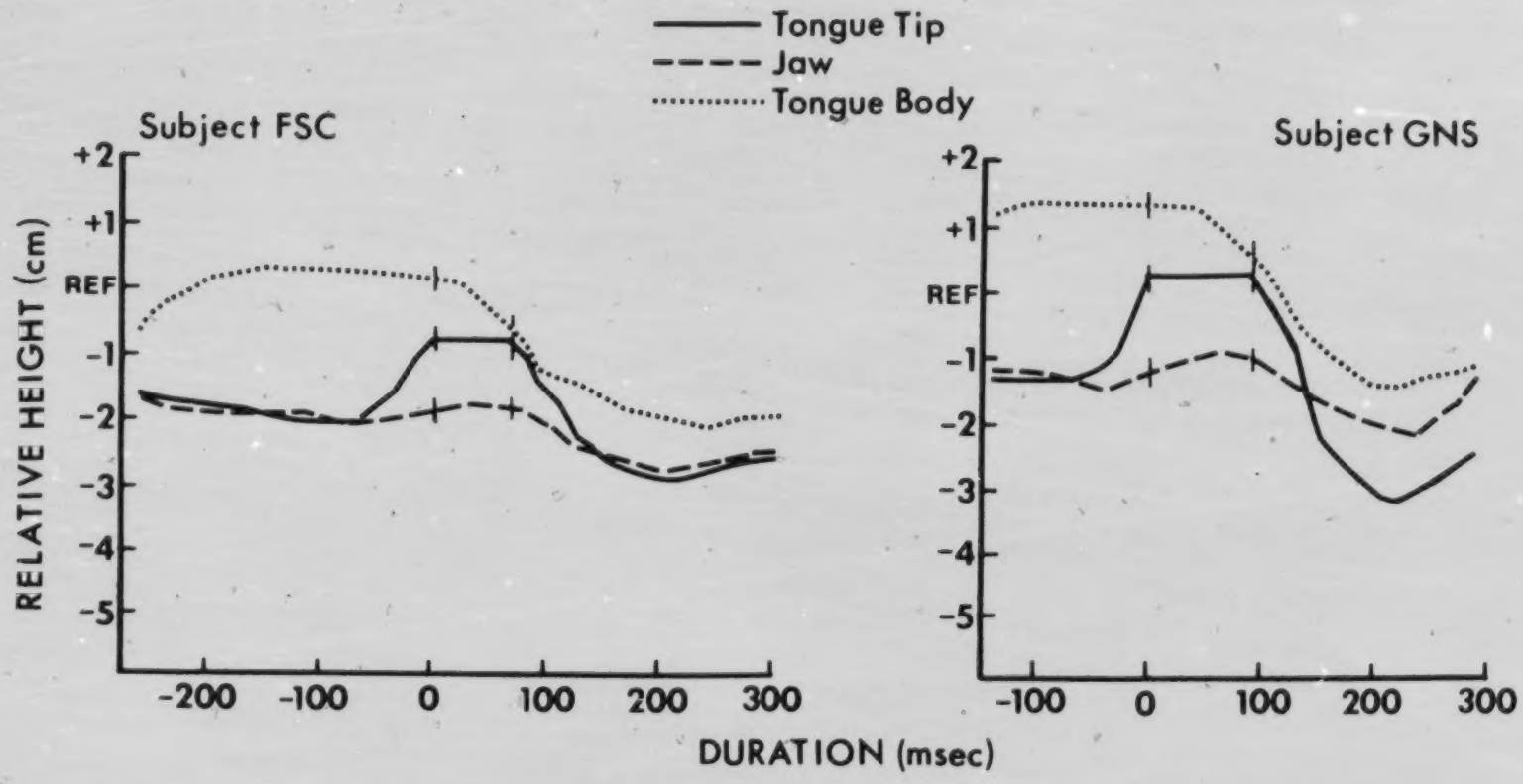


FIGURE 5

Figure 5: Movement tracks for utterance /ita/. The tongue body pellet for Subject FSC is the second, more posterior, one.

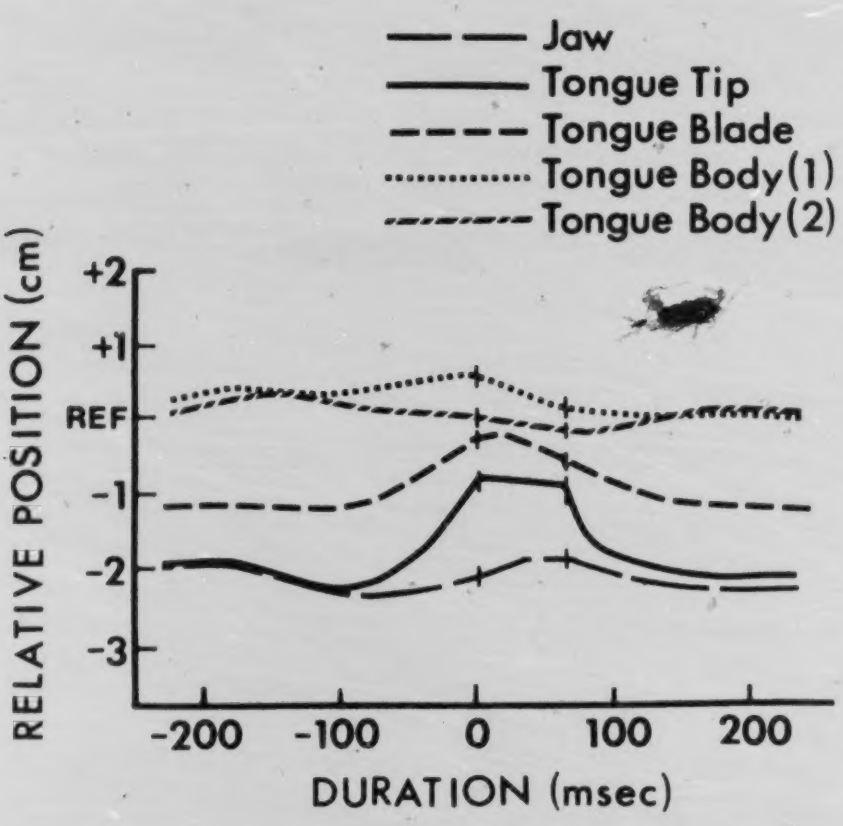


Figure 6: Movement tracks (height) for utterance /iti/, Subject FSC.

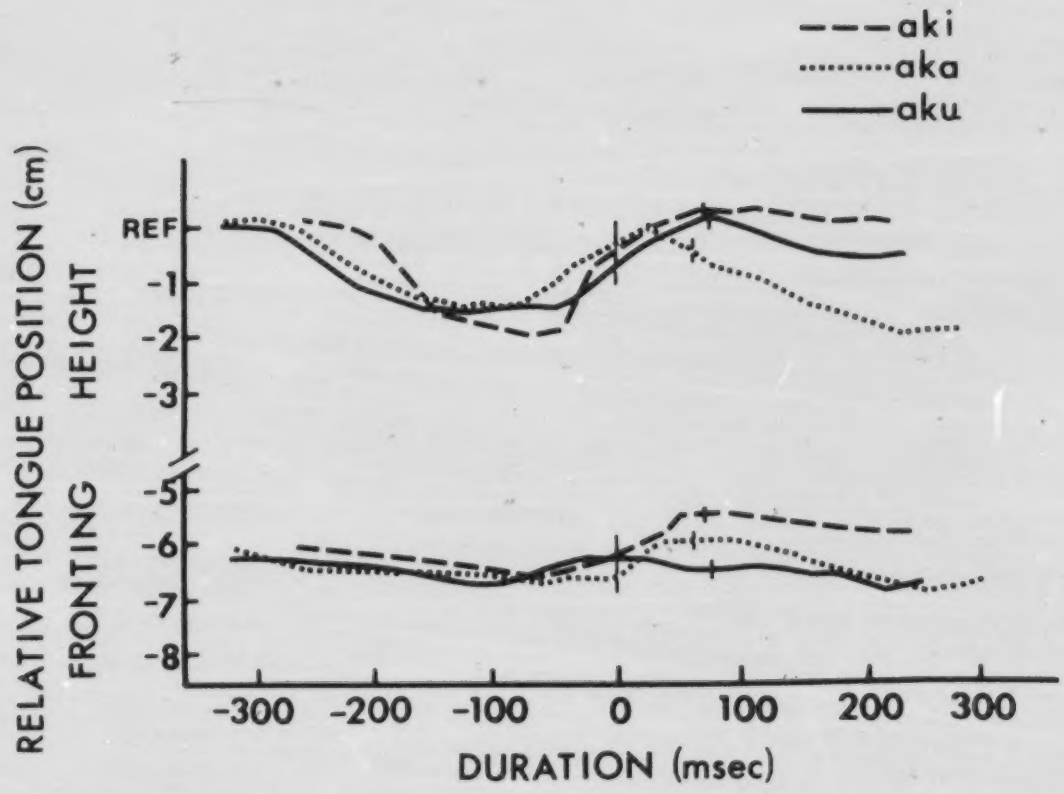


FIGURE 7

Figure 7: Movement tracks (height and fronting) for the second tongue body pellet of Subject FSC, for the utterances /aki/, /aka/, /aku/.

in the fronting dimension. Thus, consistent with the data for both /p/ and /t/, the data for /k/ show anticipatory movements to be locked to the closure phase of the consonant.

For VCV utterances containing either /p/, /t/, or /k/ as the intervocalic consonant, the usual sequence of articulatory events is as follows. Movements of the jaw, tongue body, and primary articulator begin at about the same time, with jaw closing continuing past the time of occlusion for the consonant. Shortly after closure for the consonant occurs, tongue body movement toward the second vowel begins. This movement is followed independently by jaw opening and release of the consonant. Articulatory movements for the postvocalic vowel always begin between the time of consonant closure and consonant release.

The data of this experiment, in showing consonant constraints on vowel movement in a VCV utterance, are not consistent with Öhman's (1966) hypothesis that vowel-to-vowel movement in a VCV sequence is essentially diphthongal. Öhman's hypothesis is based on the assumption that tongue body movements toward the second vowel begin at about the time of onset of closing for the consonant. However, the present data show that movement toward the second vowel begins much later, some 5 - 60 msec after closure for the consonant has already been completed. This pattern of movement even occurs for /VpV/ sequences, where the tongue body is not actively involved in the production of the intervocalic consonant. These data suggest that either the tongue body itself attains a target during consonant production, or more likely, that the release of the consonant and the movement toward the vowel are linked in a basic gesture.

In addition to questions concerning anticipatory movements of the tongue body, it was expected that the data of this experiment could be used to track the onset of lip rounding for a rounded vowel preceded by a variety of different phones. Lateral view x-rays can provide an indication of lip rounding in the form of degree of lip protrusion. Unfortunately, however, this measure was not a very sensitive one for the two speakers used in this experiment. The difference in protrusion between the spread vowel /i/ and the rounded vowel /u/ averaged only 5 mm for both speakers. It might be noted though, that in no case did evidence of a protruding gesture appear for the rounded second vowel in any of the VCV utterances until after closing for the intervocalic consonant was completed.

To summarize the data thus far: the relative timing of articulatory movements in a VCV sequence is affected by the intervocalic consonant, even if the gesture for the consonant is not a contradictory one. The intervocalic consonant affects both tongue body and jaw movements toward the second vowel. Anticipatory movements toward the second vowel always begin during the closure period of the intervocalic consonant, suggesting that the CV component of the VCV sequence might be organized as a basic unit.

The Attainment of Articulatory Targets

Carryover coarticulation effects were studied in relation to both the influence the first vowel exerts on the position of the intervocalic consonant and the influence the intervocalic consonant exerts on the attain-

ment of the target for the second vowel.

In contrast to timing measurements, useful positional measurements for /p/ could not be obtained. The important positional information for /p/ appears primarily in the coronal plane; lateral view x-rays simply do not reveal this information. However, the present data do show a rather strong vowel effect on jaw position during /p/. Figure 8 illustrates this effect for both subjects. These plots, which agree with the data of Sussman, MacNeilage, and Hanson (1973) and Gay (1974c) show that the position of the jaw during the production of /p/ is sensitive to the openness of the adjacent vowel: greater jaw opening for the consonant occurred with a more open adjacent vowel. This figure also shows what is presumed to be a stress effect in the data of Subject GNS. Jaw opening (and consequently tongue height) for /a/ is reduced in the preconsonantal position.

Carryover effects of the first vowel on the positional properties of /t/ did not appear in either the tongue tip or jaw measurements. Figure 9 illustrates the insensitivity of tongue tip position for /t/ to different preceding vowels, in both the height and fronting dimensions. It is apparent that neither the retrusiveness of /u/ nor the openness of /a/ had any measurable effect on the /t/ target, in either dimensions. The only differences in the three traces appear in the timing of the closing movements. Since the onset of closing is earliest for /a/ and latest for /u/, the differences are presumed to be displacement related. Finally, jaw movements for /t/, unlike those for /p/, were not affected by the openness of the preceding or following vowel.

The most interesting and extensive carryover effects of the first vowel on consonant production appeared in the movement track of the tongue body during /k/ production. This is illustrated in Figure 10 for the VCV sequence where /i/ is the common second vowel. Here the predicted effect of different first vowels is evident. At the time of closure for /k/, the tongue body is higher and more fronted for /i/, and progressively lower and more retruded for /u/ and /a/. The magnitude of these effects is on the order of 7 mm between /i/ and /a/ in the height dimension, and 5 mm between /i/ and /a/ in the fronting dimension. The most interesting feature of this graph, however, is that the carryover effects of the first vowel do not extend far into consonantal closure. On the contrary, the three curves converge before consonant release at about the time movement begins toward the second vowel. The relative invariance of the movement from consonant release toward the second vowel further strengthens the suggestion that the CV transition is produced as an integral unit.

Carryover effects of the first vowel on the production of the intervocalic consonant were variable: they could not be adequately measured for /p/, they did not appear for /t/, but did appear, in a predictable way, for /k/. The jaw effect evident for /p/ is apparently due to the secondary importance of jaw closure in bringing about lip closure for /p/. Although closure for /p/ can have both lower lip and jaw components, the jaw component is probably facilitory and, as such, sensitive to phonetic environment. Likewise, the difference in effects for /t/ and /k/ is presumably related to the differences in degree of involvement of the tongue body during the production of the two consonants.

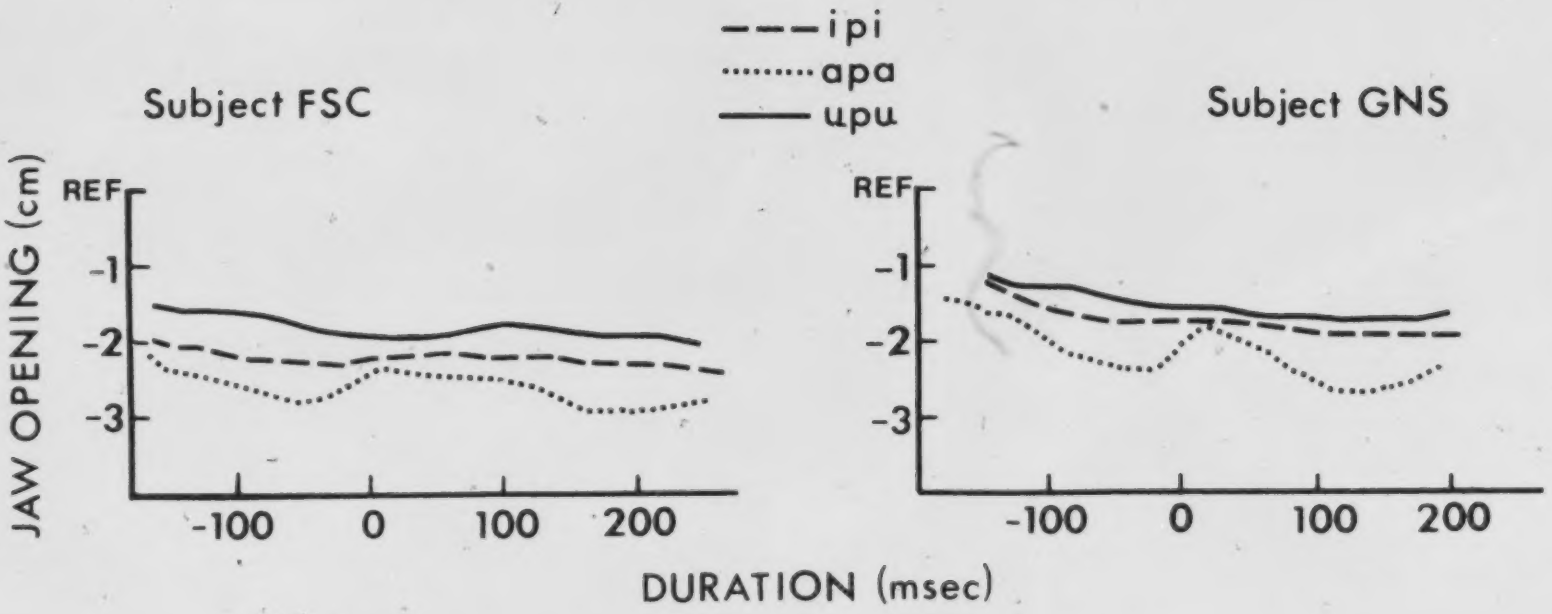


FIGURE 8

Figure 8: Movement tracks of jaw opening for /ipi/, /apa/, /upu/, both subjects.

159

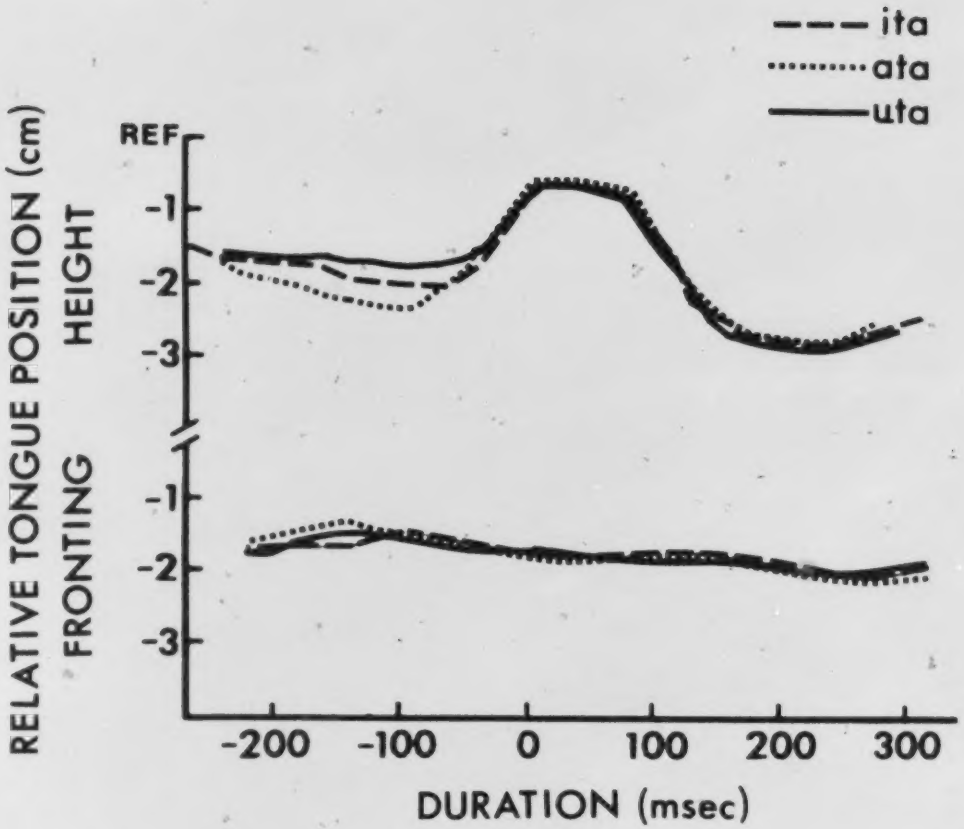


Figure 9: Movement tracks for tongue tip position, Subject FSC.

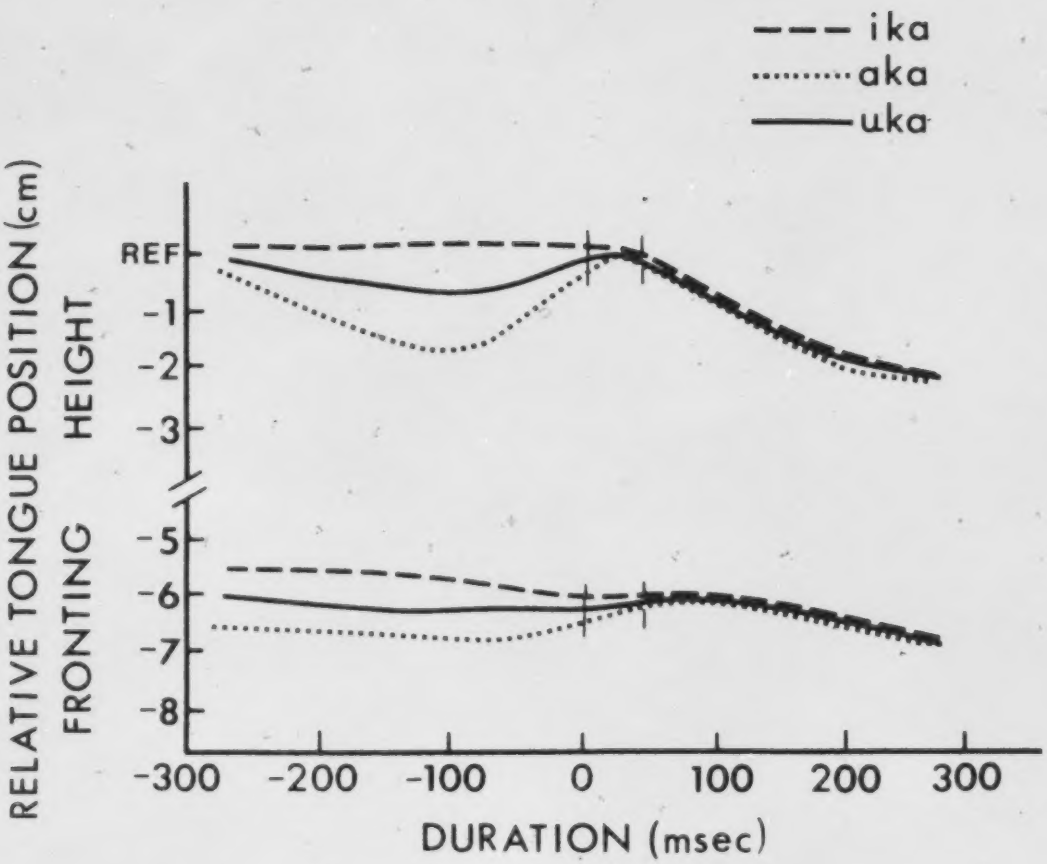


Figure 10: Movement tracks of second tongue body pellet for /ika/, /aka/, /uka/, Subject FSC.

Carryover effects of the intervocalic consonant on the following vowel appeared only for the open vowel /a/, and were reflected in differences in jaw and, consequently, tongue body height. These effects, that are consistent with those reported by Gay (1974a), are illustrated in Figure 11. This figure shows the differences in tongue body and jaw height for the vowel /a/ when the intervocalic consonant varies from /p/ to /t/. Opening for the vowel is greater when the intervocalic consonant is /p/, as opposed to /t/. The difference in tongue body height for the first vowel is probably due to differences in stress between the two utterances. However, this was not apparent when listening to the tapes. This figure also shows what appears to be tongue body involvement during the production of /t/. The movement track for the tongue body shows greater elevation than that for the jaw during the time of consonant production. This means that the tongue body position during consonant production is not simply being carried passively by the jaw, but rather has an active muscle component underlying it as well. Although variability in tongue body and jaw opening appeared in the articulatory data for both subjects, similar variability was not reflected in the corresponding acoustic measures. Apparently, the differences in jaw position as measured anteriorly at the incisors either do not correspond to the size of the pharyngeal constriction for /a/, or are much less when the arc of rotation is measured closer to the hinger axis of the jaw.

Carryover effects of a preceding consonant on the production of the vowels /i/ and /u/ were small. These effects are summarized in Figure 12 and Table 1. The figure shows the relative positions of the upper lip, lower lip, jaw, and tongue body at the time the tongue body reached its target (point of maximum displacement) for each of nine utterances containing the vowel /i/ in final position. Table 1 shows the corresponding values of the first and second formant frequencies at that point in time.

TABLE 1: First and second formant frequency values (Hz) for the vowel /i/ in nine different VCV utterances. Each utterance number corresponds to that of Figure 12.

Utterance	Subject FSC		Subject GNS	
	F_1	F_2	F_1	F_2
1. ipi	340	2200	310	2230
2. api	360	2030	320	2250
3. upi	360	2220	300	2160
4. iti	360	2220	330	2200
5. ati	320	2120	340	2210
6. uti	350	1990	320	2120
7. iki	320	2210	320	2270
8. aki	360	2160	320	2160
9. uki	350	2190	320	2250

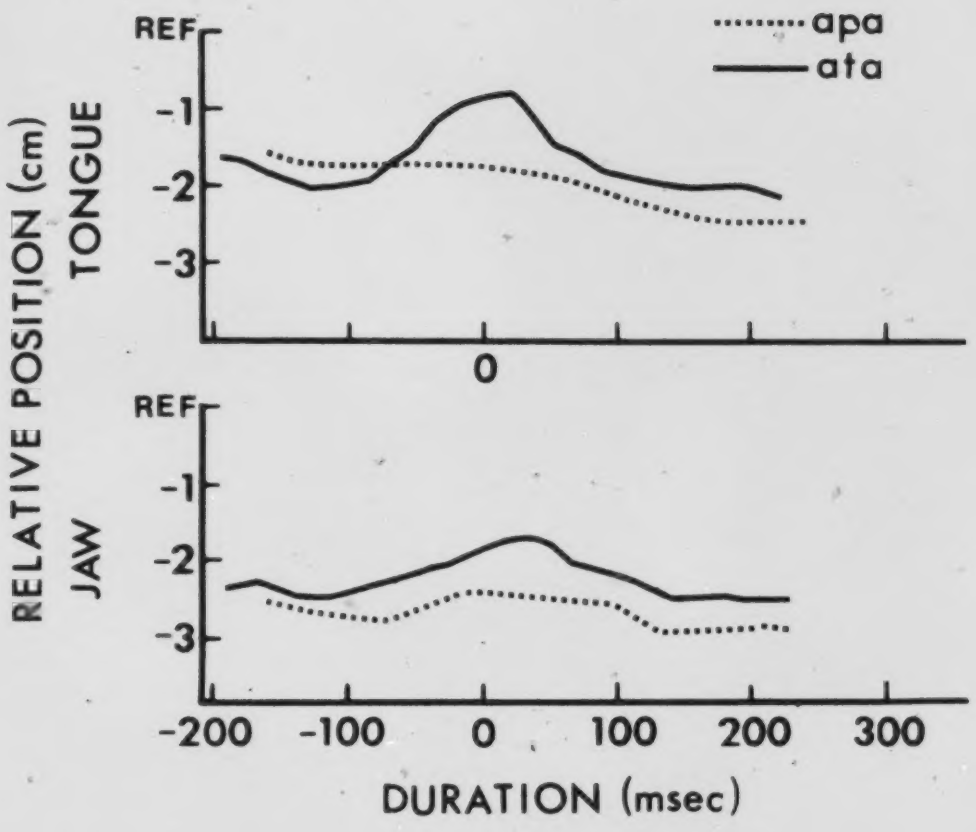


Figure 11: Movement tracks of tongue body (pellet 2) and jaw height for /apa/ and /ata/, Subject FSC.

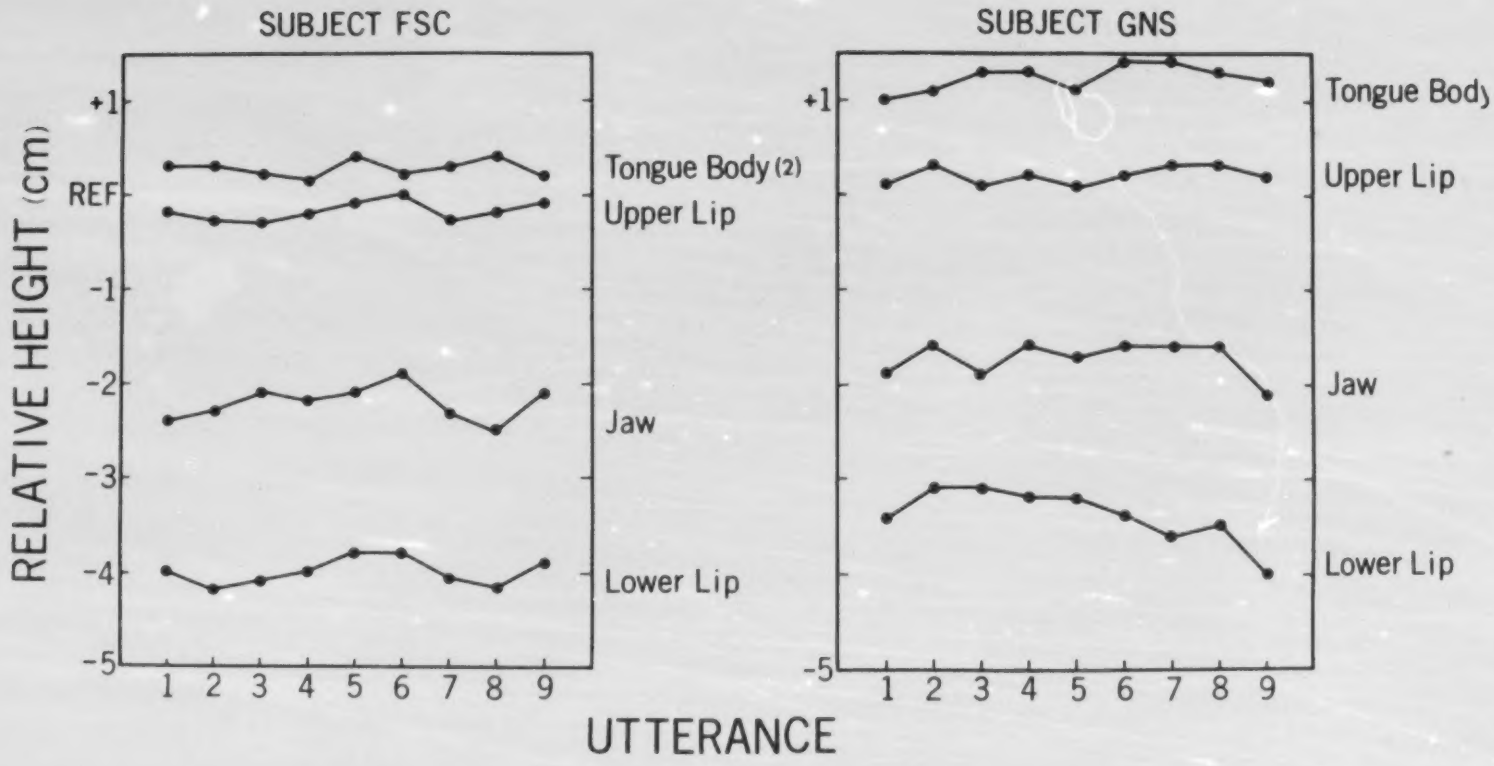


FIGURE 12

Figure 12: Coordinate positions of upper lip, lower lip, jaw, and tongue body for the target positions of the vowel /i/, both subjects. Each utterance number corresponds to the utterance in Table I.

As is evident in the figure, variability of tongue body target positions is minimal (2.5 mm for Subject FSC and 3 mm for Subject GNS). Lower lip and jaw positions, on the other hand, vary within a larger range, approximately 5 mm for Subject FSC and 10 mm for Subject GNS. Interestingly, lower lip and jaw targets seem to vary independently from tongue body positions, but covary for both subjects. This finding contradicts that of Hughes and Abbs (1976), who showed that mouth opening for /i/ remained relatively constant because of trade-offs between lower lip and jaw displacements. This type of equivalence was not evident in the present data for either /i/ or /t/. Differences between the two sets of data might be attributable to differences in either or both the speech material and instrumental methods used in the two experiments.

The acoustic measurements of target formant frequencies showed some variability among the nine utterances (Table 1). First formant frequencies were within a range of 40 Hz for both subjects, while second formant frequencies fell within a range of 230 Hz for Subject FSC, and 120 Hz for Subject GNS. The measured acoustic variability did not appear to correspond to any observed articulatory variability. For example, utterances 2 and 7 for Subject FSC were characterized by similar articulatory target points, but quite different formant frequencies. Conversely, utterances 3 and 4, and 1 and 9, were characterized by virtually the same formant frequencies, but different articulatory target points. Either the variability observed fell, for the most part, within the range of measurement error, or more likely, a four-point parameterization tracking procedure of the type used in this experiment is simply inadequate for the purpose of relating differences in articulatory target points to the acoustic output. It might also be noted that acoustic variability for both /u/ and /a/ were, in terms of percentage, within the same range as variability for /i/.

Carryover effects, then, when they do appear, are unlike anticipatory effects in that they depend on the phonetic identity of the particular segment. Like anticipatory effects, however, carryover effects seem to spread no farther than the neighboring phone. These findings support an articulatory based formulation of speech production (MacNeilage, 1970). For the most part, an articulatory target corresponded as a relatively invariant representation of a phoneme. Articulatory variability, when it did occur, did so only under special circumstances. First, carryover effects for a consonant are reflected mainly in variability of jaw position, and only when the jaw is not primarily involved in the production of the phone, as in /p/. However, when the jaw is more tightly involved in the production of a phone (/t/ for example), degree of jaw opening was not sensitive to that of the adjacent phone. The only other strong carryover effect appeared in tongue body movements for intervocalic /k/. Here, unlike variability in jaw opening, carryover effects on tongue body movements do not seem to be either random in appearance or inertial in origin. Unlike /VpV/ and /VtV/ sequences where the tongue body is usually in a waiting position before it moves toward the second vowel during consonant closure in a /VkV/ sequence, the tongue body is involved as a primary articulator in the production of the consonant. The movements of the tongue body through /k/ (Figure 10) seem to be directed, in a straight-line fashion, to a common target position for release of the consonant. The data for /k/ provide a fairly convincing illustration of the limited spreading effects of coarticulation in a VCV sequence. Because of

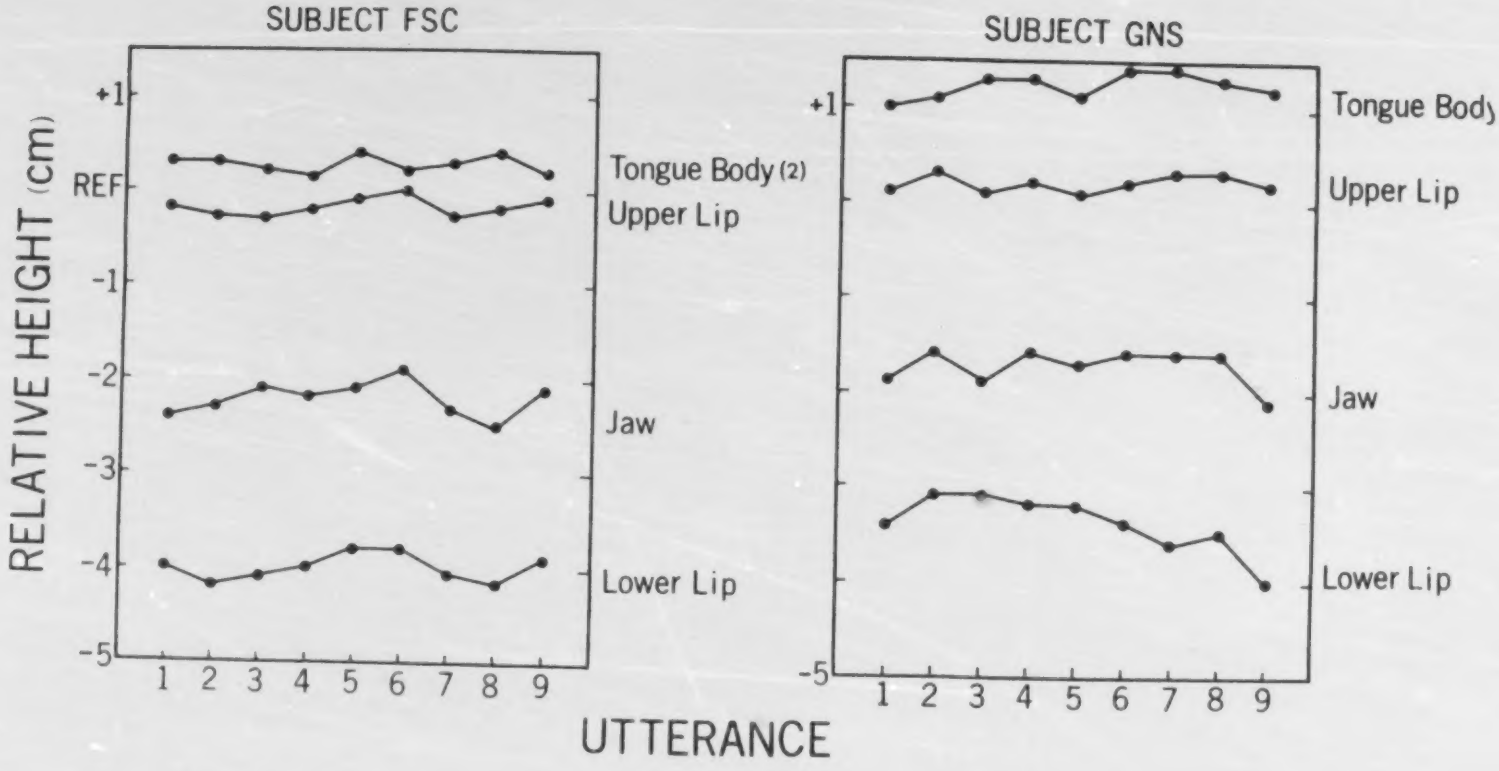


FIGURE 12

Figure 12: Coordinate positions of upper lip, lower lip, jaw, and tongue body for the target positions of the vowel /i/, both subjects. Each utterance number corresponds to the utterance in Table I.

As is evident in the figure, variability of tongue body target positions is minimal (2.5 mm for Subject FSC and 3 mm for Subject GNS). Lower lip and jaw positions, on the other hand, vary within a larger range, approximately 5 mm for Subject FSC and 10 mm for Subject GNS. Interestingly, lower lip and jaw targets seem to vary independently from tongue body positions, but covary for both subjects. This finding contradicts that of Hughes and Abbs (1976), who showed that mouth opening for /i/ remained relatively constant because of trade-offs between lower lip and jaw displacements. This type of equivalence was not evident in the present data for either /i/ or /t/. Differences between the two sets of data might be attributable to differences in either or both the speech material and instrumental methods used in the two experiments.

The acoustic measurements of target formant frequencies showed some variability among the nine utterances (Table 1). First formant frequencies were within a range of 40 Hz for both subjects, while second formant frequencies fell within a range of 230 Hz for Subject FSC, and 120 Hz for Subject GNS. The measured acoustic variability did not appear to correspond to any observed articulatory variability. For example, utterances 2 and 7 for Subject FSC were characterized by similar articulatory target points, but quite different formant frequencies. Conversely, utterances 3 and 4, and 1 and 9, were characterized by virtually the same formant frequencies, but different articulatory target points. Either the variability observed fell, for the most part, within the range of measurement error, or more likely, a four-point parameterization tracking procedure of the type used in this experiment is simply inadequate for the purpose of relating differences in articulatory target points to the acoustic output. It might also be noted that acoustic variability for both /u/ and /a/ were, in terms of percentage, within the same range as variability for /i/.

Carryover effects, then, when they do appear, are unlike anticipatory effects in that they depend on the phonetic identity of the particular segment. Like anticipatory effects, however, carryover effects seem to spread no farther than the neighboring phone. These findings support an articulatory based formulation of speech production (MacNeilage, 1970). For the most part, an articulatory target corresponded as a relatively invariant representation of a phoneme. Articulatory variability, when it did occur, did so only under special circumstances. First, carryover effects for a consonant are reflected mainly in variability of jaw position, and only when the jaw is not primarily involved in the production of the phone, as in /p/. However, when the jaw is more tightly involved in the production of a phone (/t/ for example), degree of jaw opening was not sensitive to that of the adjacent phone. The only other strong carryover effect appeared in tongue body movements for intervocalic /k/. Here, unlike variability in jaw opening, carryover effects on tongue body movements do not seem to be either random in appearance or inertial in origin. Unlike /VpV/ and /VtV/ sequences where the tongue body is usually in a waiting position before it moves toward the second vowel during consonant closure in a /VkV/ sequence, the tongue body is involved as a primary articulator in the production of the consonant. The movements of the tongue body through /k/ (Figure 10) seem to be directed, in a straight-line fashion, to a common target position for release of the consonant. The data for /k/ provide a fairly convincing illustration of the limited spreading effects of coarticulation in a VCV sequence. Because of

continuous tongue body involvement in the production of CVC syllables containing /k/ as the intervocalic consonant, the elements of these syllables, specially /k/ itself, should be the most sensitive to the spreading of coarticulation effects in both directions. Yet, the assimilation of carryover effects and the onset of anticipatory movements both occur within the closure period of the consonant, with movements from the same vowel into /k/ (ref. Figure 7), or movements toward the same vowel from /k/ (ref. Figure 10), not being affected by the articulatory event on the other side of the consonant.

Stability of tongue body targets for vowels (at least /i/ and /u/) was also the rule rather than the exception. The only substantial articulatory variability occurred in jaw displacement, with /a/ showing the greatest effects and /u/ the least. As was mentioned before, however, variability in jaw displacement for /a/, as measured anteriorly at the incisors, might be either exaggerated or irrelevant in relation to variability that might exist in the pharyngeal constriction for /a/. Likewise, the variability of maximum jaw displacement for both /i/ and /u/ seems unrelated to the variability observed in the position of the tongue body for those vowels. Thus, the two features, tongue body height and jaw displacement, might be independent ones, with jaw opening being a facilitory gesture and an unmarked phonetic feature. This formulation suggests a reevaluation of models of vowel articulation that specify jaw position as a primary determiner of tongue height (Lindblom and Sundberg, 1971).

SUMMARY AND CONCLUSIONS

The major findings produced by this experiment are as follows. First, anticipatory movements toward the second vowel in a vowel-stop consonant-vowel sequence begin during the closure period of the intervocalic consonant. This restricted coarticulatory field includes both tongue body and jaw movements associated with the second vowel. Furthermore, the size of this field is not affected by the identity of the intervocalic consonant. Second, like anticipatory effects, carryover effects did not extend beyond an immediately neighboring segment. Unlike anticipatory effects, however, the appearance of carryover coarticulation effects depended on the phonetic identity of the particular segment on which these effects might act.

The implication of these findings is that the rules governing the segmental input to a VCV string might not be as complex as present models suggest. The finding that anticipatory movements begin and primary carryover effects end at about the same time during the closure period of the consonant, suggests that the release of the consonant and movement toward the vowel are organized and produced as an integral articulatory event.

This formulation, which specifies a syllable-sized articulatory unit, is not consistent with the operation of a phoneme based scan-ahead mechanism. This does not necessarily mean, however, that a scan-ahead mechanism does not operate on larger units or at another stage of the speech production process. For example, Lindblom and Rapp (1973), Nooteboom and Cohen (1975), and Fromkin (1971) have suggested the existence of an anticipatory mechanism in the temporal formulation of speech sequences. Likewise, the complex reordering of commands accompanying changes in speaking rate (Gay, Ushijima, Hirose,

and Cooper, 1974) also suggests that the temporal features of a downstream segment might be known in advance.

Thus, while it has traditionally been considered that the serial ordering of segments is governed by complex rules whose effects can spread across several adjacent segments, and the temporal control of speech is governed by a simple adjustment of timing of commands to the articulators (Lindblom, 1973), it may well be that the reverse is true: the segmental input to the speech string is governed primarily by simple rules that act upon syllable-sized units, while the temporal formulation of the string requires complex articulatory adjustments based on advance information obtained from a higher level scan-ahead mechanism.

Like most studies of speech organization, especially those using high-speed cinefluorographic techniques, the results of this experiment are based on data obtained from a relatively small subject population and are applicable to the production of only a few phonetic elements, themselves constrained by the artificial format in which they were placed. Thus, the findings of this experiment are obviously far from conclusive, and go only part way toward answering those questions posed at the outset. The present findings can serve, however, as a basis for examining or reexamining a number of questions concerning the organization of segmental gestures. For example, it was shown that a four-point parameterization procedure for relating articulatory targets to acoustic targets is inadequate. In order to resolve the differences between the acoustic data of Öhman (1966) and the articulatory data of the present study, formant tracking must be matched to a far more comprehensive multipoint parameterization of the vocal tract. The present results also suggest, without providing convincing evidence, that the onset of anticipatory lip rounding might be conditioned differently in CCCV and VCV sequences; also, they raise further questions about the use of trade-offs between tongue and jaw movements in achieving articulatory targets, and the importance of jaw position in determining tongue height in vowel articulation.

REFERENCES

Bell-Berti, F. and K. S. Harris. (1975) Some acoustic measures of anticipatory and carryover coarticulation. Haskins Laboratories Status Report on Speech Research SR-42/43, 297-304.

Benguerel, A-P. and H. A. Cowan. (1974) Coarticulation of upper lip protrusion in French. Phonetica 30, 41-55.

Daniloff, R. G. and K. L. Moll. (1968) Coarticulation of lip-rounding. J. Speech Hearing Res. 11, 707-721.

Fromkin, V. A. (1971) The non-anomalous nature of anomalous utterances. Language 47, 27-52.

Gay, T. J. (1974a) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.

Gay, T. J. (1974b) Some electromyographic measures of coarticulation in VCV utterances. Haskins Laboratories Status Report on Speech Research SR-44, 137-145.

Gay, T. J. (1974c) Jaw movements during speech: A cinefluorographic investigation. Haskins Laboratories Status Report on Speech Research SR-39/40, 219-230.

Gay, T., T. Ushijima, H. Hirose, and F. S. Cooper. (1974) Effect of speaking rate on labial consonant-vowel articulation. J. Phonetics 2, 46-63.

Henke, W. (1966) Dynamic Articulatory Model of Speech Production Using Computer Simulation. Ph.D. Thesis, M.I.T.

Houde, R. A. (1967) A Study of Tongue Body Motion During Selected Speech Sounds. Ph.D. Thesis, University of Michigan.

Hughes, O. M. and J. H. Abbs. (1976) Labial-mandibular coordination in the production of speech: Implications for the operation of motor equivalence. Phonetica 33, 199-201.

Kent, R. D. (1970) A cinefluorographic-spectrographic investigation of the component gestures in lingual articulation. (Ph.D. Thesis, University of Iowa).

Kozhevnikov, V. A. and L. A. Chistovich. (1965) Rech', Artikulyatsiya, i Vospriyatiye. Trans. as Speech: Articulation and Perception. (Washington, D.C.: Joint Publications Research Service) 30, 543.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B. and K. Rapp. (1973) Some temporal regularities of spoken Swedish. PILUS, (Stockholm Univ.).

Lindblom, B. E. F. and J. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw and larynx movement. J. Acoust. Soc. Am. 50, 1166-1179.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-195.

MacNeilage, P. F. and J. DeClerk. (1969) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Am. 45, 1217-1233.

McClean, M. (1973) Forward coarticulation of velar movement at marked junctural boundaries. J. Speech Hearing Res. 16, 286-296.

Moll, K. L. and R. G. Daniloff. (1971) Investigation of the timing of velar movements during speech. J. Acoust. Soc. Am. 50, 678-684.

Nooteboom, S. G. and A. Cohen. (1975) Anticipation in speech and its implications for perception. Proceeding of the Symposium on Dynamic Aspects of Speech Perception (IPR, Eindhoven).

Ohde, R. N. and D. J. Sharf. (1974) Coarticulatory effects of voiced stops on the reduction of acoustic vowel targets. J. Acoust. Soc. Am. 58, 923-924.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge, Mass.; MIT Press).

Sussman, H., P. F. MacNeilage, and R. Hanson. (1973) Labial and mandibular dynamics during the production of bilabial consonants. J. Speech Hearing Res. 16, 397-420.

Ushijima, T. and H. Hirose. (1975) Velar movement and its motor command. Haskins Laboratories Status Report on Speech Research SR-41, 207-216.

Measuring Laterality Effects in Dichotic Listening*

Bruno H. Repp

ABSTRACT

This paper discusses methodological issues and problems related to measuring laterality effects in dichotic listening. Section 1 describes the standard dichotic two-response paradigm as well as a number of indices of the ear advantage proposed in the literature. The numerical range of most of these indices is constrained by performance level; only one particular index avoids these constraints. However, this does not necessarily make this index the optimal one. A correction for guessing is proposed--an issue that has been neglected in the past. Analogies to signal detection theory are discussed, as well as the theoretical and empirical criteria for choosing the "correct" index of laterality. The index called e_g is proposed as the best solution given the present state of knowledge. Section 2 discusses the phenomenon of dichotic fusion and the dichotic single-response paradigm, which offers many methodological advantages over the two-response paradigm. Section 3 discusses the factors of ear dominance and stimulus dominance in the perception of fused stimuli. An index of ear dominance is derived by taking advantage of analogies to signal detection theory. In Section 4, a number of remaining problems are discussed: stimulus intelligibility, guessing and selective attention, blend responses, test reliability, validity, and homogeneity.

INTRODUCTION

Since Kimura's (1961) demonstration of an average right-ear advantage (REA) in the recognition of dichotic verbal stimuli, many researchers have used dichotic listening tasks to measure hemispheric dominance for language. Kimura's interpretation that hemispheric dominance for language underlies the ear asymmetries has had almost universal acceptance. While some studies have been content with diagnosing the mere direction of the average ear advantage (left or right) and testing its significance, many recent studies have

---

*A slightly revised version of this paper is now in press in the Journal of the Acoustical Society of America.

Acknowledgment: I would like to thank Terry Halwes for numerous discussions and comments in all stages of this project. In the early stages, this work was supported by NIH Grant DE00202 to the University of Connecticut Health Center, Farmington. Later, support came from NICHD Grant HD01994 to the Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

attempted to compare different individuals, different tests, or different experimental conditions with respect to the observed magnitude of the ear asymmetry. Underlying these attempts has been the belief that cerebral lateralization, like handedness, is a matter of degree and can be measured on a continuous scale (Zangwill, 1960; Shankweiler and Studdert-Kennedy, 1975).

In order to yield meaningful and reliable measurements, dichotic testing must meet certain formal and methodological requirements that have been given relatively little attention in the past. If dichotic listening tasks are used as instruments to measure the degree of hemispheric dominance for language, they must satisfy the same high standards of construction, procedure, and scoring as any other psychological test. These standards may be derived from methodologically oriented research in the laboratory, theoretical analyses of the task situation, and general test-theoretical principles. Many of these requirements are not sufficiently met by dichotic tests as they are now used.

The present paper summarizes the issues that must be handled in constructing a good dichotic test to measure hemispheric dominance. The dichotic listening situation is remarkably complex. In the discussion that follows, I provide some suggestions, but point out many problems that need further investigation or have not been dealt with at all in the past. Although the discussion is restricted to dichotic listening, many of the issues should apply to any situation in which lateral asymmetries are to be measured (for example, tachistoscopic perception, binocular rivalry, or ocular dominance experiments), and therefore may be of interest to a wider audience.

The first focus of the present discussion is choosing a numerical index of the ear advantage. This problem is fundamental to the measurement of lateralization; unless it is solved, no meaningful comparisons between subjects, tests, or experimental conditions are possible. In Section 1--which heavily relies on earlier discussions by Halwes (1969) and Marshall, Caplan, and Holmes (1975)--I discuss a number of indices that have been proposed and used in the past in conjunction with the dichotic two-response paradigm (that requires the listener to identify both stimuli in a dichotic pair). Most of these indices fail to take into account the constraints imposed by performance level on the range of differences between the scores for the two ears. In addition, none of them corrects for guessing, despite the fact that most dichotic studies use only a few different stimuli, resulting in substantial guessing probabilities. After describing an index that takes both performance level and guessing into account, I hasten to point out that a correct index must be based on a correct theory and empirical evidence of how scores for the two ears change with performance level and how guessing operates. This theoretical and empirical basis is not available at present. I describe an index that is based on plausible assumptions, but the question whether it is <u>the</u> "correct" index remains open.

The second focus of the present paper is finding ways to simplify dichotic testing and to circumvent some of the problems encountered in the standard two-response paradigm. In Sections 2 and 3, I discuss an approach to dichotic listening that in many ways seems simpler than the two-response paradigm. This method, that requires only a single response to each dichotic

stimulus, relies on the phenomenon of dichotic (or binaural) fusion. In Section 2, I discuss the factors that make two dichotic stimuli fuse more or less completely into a single perceived stimulus, as well as the methodological consequences of such fusion. Section 3 derives an index of the ear advantage for the single-response paradigm. In the course of deriving the index, I discuss the phenomenon of stimulus dominance (perceptual dominance of one stimulus over the other in a fused dichotic pair) that exerts constraints on the ear score difference similar to that exerted by performance level in the two-response paradigm. I illustrate how these constraints can be dealt with and how they actually become a crucial factor in deriving an unbiased index of the ear advantage.

Section 4 is devoted to a survey of additional topics and problems in dichotic testing: stimulus intelligibility, selective attention, blend responses, test reliability, homogeneity, and validity. Since my concern in this paper is exclusively methodological, I avoid any discussion of the physiological factors that may underly dichotic ear advantages. My aim is to develop methods for measuring the dichotic ear advantage with maximum precision. Before we can attempt to answer the more fundamental questions about the structures and processes underlying the ear asymmetry, we must be able to obtain valid and reliable measurements from dichotic tests. There is much room for improvement in existing methods with respect to that goal.

1. LATERALITY INDICES IN THE TWO-RESPONSE PARADIGM

1.1. The Method

In the two-response paradigm, two different stimuli are simultaneously presented to the two ears, and the subject is asked to identify both--typically without any constraint on the order of report. The two responses must be different from each other, and guessing is encouraged. This is the standard situation that will be considered in this section.

The results of a standard two-response test may be summarized in a 2 x 2 table, as shown in Table 1. The responses are scored as correct (that is, identical with one of the stimuli) or incorrect, without regard to order. The proportions of correct and incorrect responses are calculated separately for each ear, so that the row sums in Table 1 are equal to 1.0.

TABLE 1: The data structure in the two-response paradigm.

		Responses		
		Correct	Incorrect	
Channels	LE	P_L	$1 - P_L$	1.00
	RE	P_R	$1 - P_R$	1.00
		$P_L + P_R$	$2 - P_L - P_R$	2.00

The overall performance level is defined as the average proportion of correct responses per ear,

$$(1) \qquad P_o = (P_R + P_L)/2 \ .$$

1.2. The Simple Difference Score (d)

The simplest index of the ear advantage is the difference between the proportions of correct responses for the two ears,

$$(2) \qquad d = P_R - P_L \ .$$

The vast majority of dichotic listening studies have reported the ear advantage as d. (There is no commonly accepted name of the index; I call it d here simply for notational convenience.) The symbol d is useful as a descriptive statistic, but it has severe limitations when the results of different subjects, different tests, or different experimental conditions are to be compared. These limitations arise from the constraint imposed on differences between proportions by their absolute size--a fact that is often neglected and so constitutes one of the primary fallacies of descriptive statistics. In the context of measuring laterality effects, Halwes (1969) was the first to point out that the overall performance level P_o sets an upper limit to d,

$$(3) \qquad \begin{aligned} d_{max} &= 2P_o && \text{if } 0.0 \leq P_o \leq 0.5 \\ d_{max} &= 2(1 - P_o) && \text{if } 0.5 \leq P_o \leq 1.0 \ , \end{aligned}$$

where d_{max} is the maximal value that d can assume at a given level of P_o, and $d_{max} = -d_{min}$, the corresponding minimal value. Figure 1a shows the triangular function represented by Equation 3.

Thus, d indices of different subjects, tests, or experimental conditions are not directly comparable unless the respective performance levels are equal. Since, in general, performance levels are not constant from one subject (test, condition) to another, comparisons of d indices are almost certainly invalid. Many studies in the past have neglected this quite elementary limitation of simple difference scores and, consequently, some of these studies may have reached faulty conclusions.[1] I should point out that,

---

[1]Consider, for example, two subjects, A and B, with $P_o = 0.6$ and 0.8, respectively. Assume that d = 0.5 for A and d = 0.4 for B. Who shows the larger ear advantage? From a comparison of d indices, the answer would be A. However, B never could have reached that index because of her higher performance level that permits a maximal d of only 0.4. There is no reason why B's better performance on the test should imply that she is less lateralized than A. In fact, once performance level is taken into account, it becomes clear that B shows the maximal d for her level of performance, while A's index is considerably below the maximal d possible at $P_o = 0.6$ ($d_{max} = 0.8$). It therefore should be concluded that, contrary to the first superficial impression, B shows a stronger REA than A.

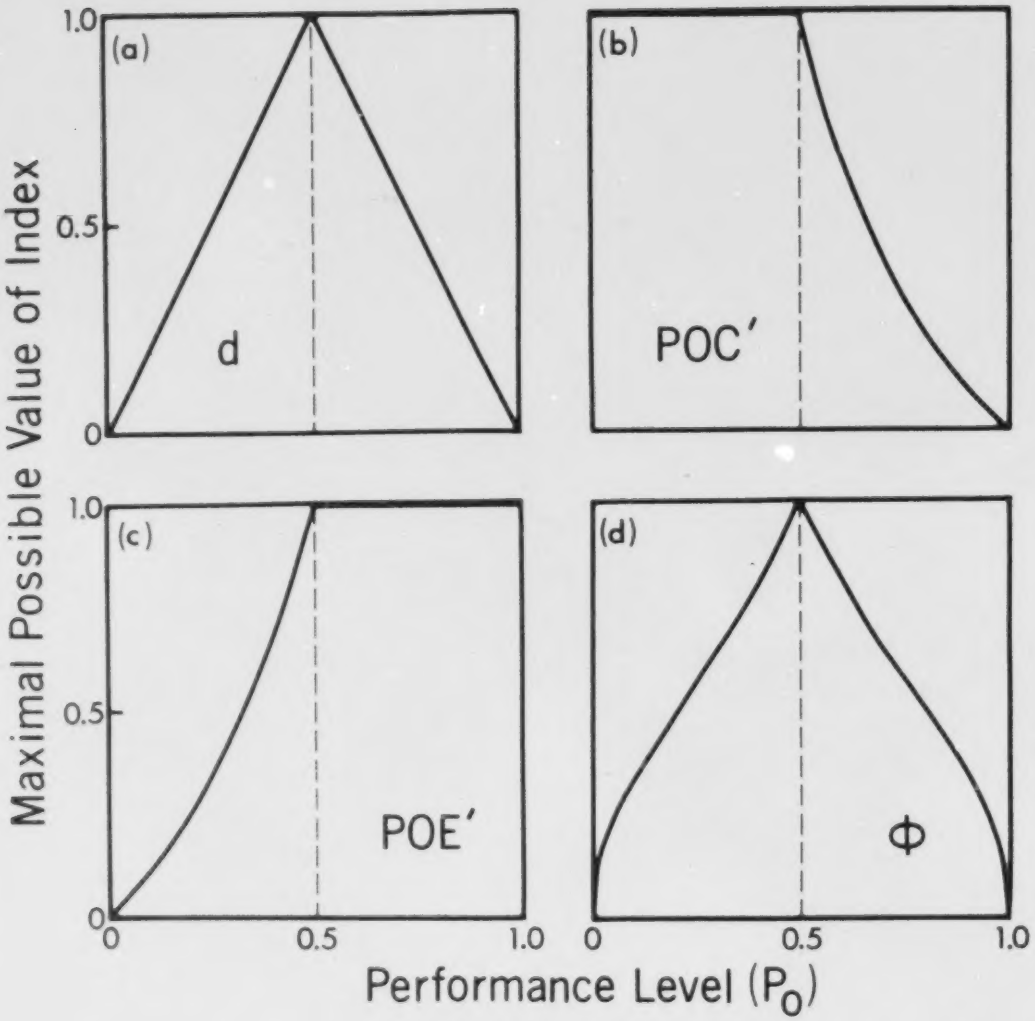


Figure 1: The numerical ranges of four indices of the ear advantage as a function of performance level.

theoretically, it does remain conceivable that direct comparisons of d are valid, after all--that is, that d is the correct index to use--but the assumptions that would have to be made in order to justify such comparisons are highly implausible (see Section 1.5). Hopefully, empirical evidence will become available in the future to decide this issue objectively.

1.3. "Correcting" for the Constraints of Performance Level

Several authors became aware of the limitations of d and proposed alternative indices of the ear advantage (that were subsequently used by others). All of these indices were intended to provide a measure of the ear advantage that is independent of performance level, both theoretically and empirically. Only the last of the four indices that I will discuss seems to achieve this aim, but, as I will argue, it is still far from a final solution to the problem of finding the optimal index of the ear advantage.

POC (percentage of correct [responses]) and POE (percentage of errors) are two alternative indices suggested by Harshman and Krashen (1972). They are defined as

$$(4) \qquad POC = P_R/(P_R + P_L) \ ,$$

$$(5) \qquad POE = (1 - P_L)/(2 - P_R - P_L) \ .$$

These indices range from 0 (perfect LEA) to 1 (perfect REA); an index of 0.5 means no ear advantage. For those who, like myself, prefer a scale ranging from -1 (perfect LEA) to +1 (perfect REA)--and this is entirely a matter of personal choice--corresponding POC' and POE' indices are obtained by a simple linear transformation of POC and POE:

$$(6) \qquad POC' = 2POC - 1 = (P_R - P_L)/(P_R + P_L) \ ,$$

$$(7) \qquad POE' = 2POE - 1 = (P_R - P_L)/(2 - P_R - P_L) \ .$$

An analagous index was proposed by Studdert-Kennedy and Shankweiler (1970), but their index was based on single-correct responses only. For a brief discussion, see Repp (1977b).

The limitations of POC and POE as a function of performance level have recently been competently discussed by Marshall et al. (1975). The analogous limitations of POC' and POE' are illustrated in Figures 1b and 1c. In formal terms, we obtain from Equations 3, 6, and 7,

$$(8) \qquad \begin{aligned} POC'_{max} &= 1 && \text{if } 0.0 \leq P_o \leq 0.5 \\ POC'_{max} &= (1 - P_o)/P_o && \text{if } 0.5 \leq P_o \leq 1.0 \ , \end{aligned}$$

$$(9) \qquad \begin{aligned} POE'_{max} &= P_o/(1 - P_o) && \text{if } 0.0 \leq P_o \leq 0.5 \\ POE'_{max} &= 1 && \text{if } 0.5 \leq P_o \leq 1.0 \ . \end{aligned}$$

Thus, it is evident that the range of POC' is unconstrained at low performance levels and the range of POE' is unconstrained at high performance levels, but where one index is unconstrained the other is severely limited by performance level. The same is true for POC and POE. Harshman and Krashen

(1972) preferred POE over POC after empirically demonstrating a high positive correlation between P_o and POC, but a low correlation between P_o and POE, as computed over a number of studies in the literature. This finding can be explained by the fact that high performance levels are more commonly encountered in dichotic studies than low performance levels, so that the majority of the reported scores fell in the region where POE_{max} rather than POC_{max} is independent of performance level.

A quite different and highly original approach was taken by Kuhn (1973) who proposed an existing statistical index, the ϕ coefficient, as the solution to the performance level problem. However, Levy (in press) has presented mathematical proof and empirical evidence that the ϕ coefficient does depend on performance level. The theoretical argument can be made in simplified form by pointing out the relationship between ϕ and POC' and POE':

$$(10) \qquad \phi = (P_R - P_L)/[(P_R + P_L)(2 - P_R - P_L)]^{1/2} = [(POC')(POE')]^{1/2} .$$

Then, from Equations 8, 9, and 10,

$$(11) \qquad \begin{aligned} \phi_{max} &= [P_o/(1 - P_o)]^{1/2} \quad \text{if } 0.0 \leq P_o \leq 0.5 \\ \phi_{max} &= [(1 - P_o)/P_o]^{1/2} \quad \text{if } 0.5 \leq P_o \leq 1.0 . \end{aligned}$$

Thus, ϕ_{max}, much like d_{max}, is constrained by P_o at all performance levels except 0.5. This is illustrated in Figure 1d.

Being a conjunction of POC' and POE'--viz., their geometric mean--ϕ combines the constraints of these two indices. The most obvious solution is a disjunctive use of POC' and POE' that takes advantage of the fact that each is unconstrained in one half of the range of P_o. Thus,

$$(12) \qquad \begin{aligned} e &= POC' = (P_R - P_L)/(P_R + P_L) \quad \text{if } 0.0 \leq P_o \leq 0.5 \\ e &= POE' = (P_R - P_L)/(2 - P_R - P_L) \quad \text{if } 0.5 \leq P_o \leq 1.0 . \end{aligned}$$

Since $e = d/d_{max}$ (cf. Equations 2 and 3), $e_{max} = 1$ and thus is completely independent of P_o. The idea to express the observed ear difference as a proportion of the maximally possible ear difference at a given performance level was first conceived by Halwes (1969) and, more recently and apparently independently, by Marshall et al. (1975) who called their index f. The solution seems straightforward--it is a simple multiplicative rescaling of d to fit its restricted range.

Nevertheless, e is not necessarily the optimal index. The kind of theoretical and empirical support that is needed to determine the correct index will be discussed in Section 1.5 (see also Marshall et al., 1975). At this point, I would like to consider a more obvious shortcoming of the e index (and all other indices proposed, for that matter): its failure to correct for guessing. Strangely enough, a correction for guessing has never been considered in the past, although it is obvious that guessing plays a substantial role in most dichotic experiments. In the next section, I will propose a correction for this factor.

1.4. Correction for Guessing

In order to deal with the guessing problem, we need to consider the scores for each ear, not just their difference d, as a function of P_o. This is illustrated in Figure 2a. The diagonal line labeled $P_R = P_L$ is the case of no ear advantage ($d = 0$). In this case, $P_R = P_L = P_o$, regardless of the guessing probability. At the other extreme, consider the maximal and minimal possible ear scores, P_{Rmax} and P_{Lmin}, as a function of P_o. (We assume here, without loss of generality, that the right ear is the dominant ear; the corresponding results for left-ear advantages are obtained by interchanging the R and L subscripts.) Let us first assume that N, the number of stimuli, equals infinity, so that the guessing probability is zero. Then the lowest possible performance level, P_{omin}, is zero, and, of course, $P_{Rmax} = P_{Lmin} = 0$ if $P_{omin} = 0$. As P_o increases, P_{Rmax} increases linearly towards 1.0 while P_{Lmin} remains at 0; consequently, $P_{Rmax} = 2P_o$ and $P_{Lmin} = 0$ for $0.0 \leq P_o \leq 0.5$. At $P_o = 0.5$, P_{Rmax} reaches 1.0 and remains at this level while P_{Lmin} begins to increase with P_o; consequently, $P_{Rmax} = 1.0$ and $P_{Lmin} = 2P_o - 1$ for $0.5 \leq P_o \leq 1.0$. Thus, the maximally divergent scores for the two ears are represented by the large parallelogram labeled $N = \infty$ in Figure 2a. Of course, $P_{Rmax} - P_{Lmin} = d_{max}$, whose relation to P_o is shown in Figure 1a and again in Figure 2b as the function labeled $N = \infty$.

Now consider the more realistic case of a nonzero guessing probability. Two typical cases, $N = 6$ and $N = 4$, are illustrated in Figure 2a. The lowest expected performance level for a given number of stimuli, P_{omin}, is found to be

$$P_{omin} = (N - 1)/\binom{N}{2} = 2/N \quad . \tag{13}$$

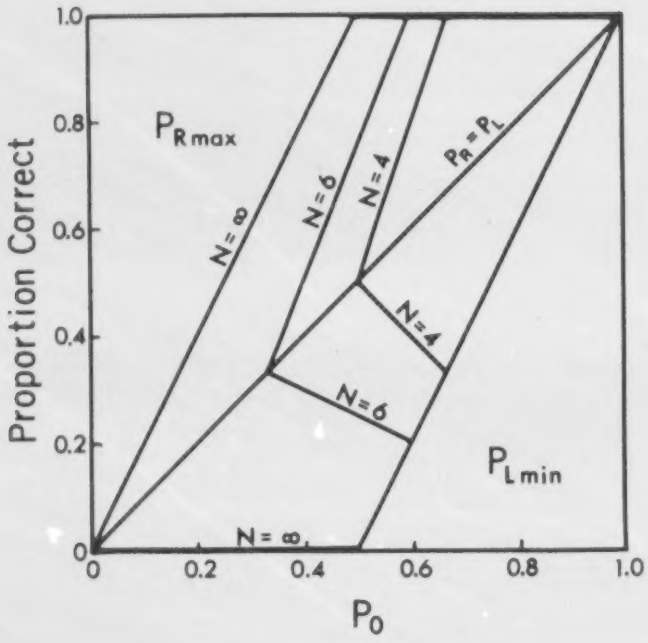
This is the performance level that would be expected if the subject produced only completely random guesses, because $(N - 1)$ of the possible $\binom{N}{2} = N(N - 1)/2$ combinations of two responses lead by chance to a correct response for one ear. Thus, $P_{Rmax} = P_{Lmin} = P_{omin}$ if $P_o = P_{omin}$. From this minimum, P_{Rmax} increases linearly towards 1.0 as P_o increases, while P_{Lmin} remains at chance level. However, this chance level does not remain constant but depends on P_{Rmax}. At the point of maximal ear difference, P_{Rmax} reaches 1.0, and

$$P_{Lmin} = 1/(N - 1) \quad , \tag{14}$$

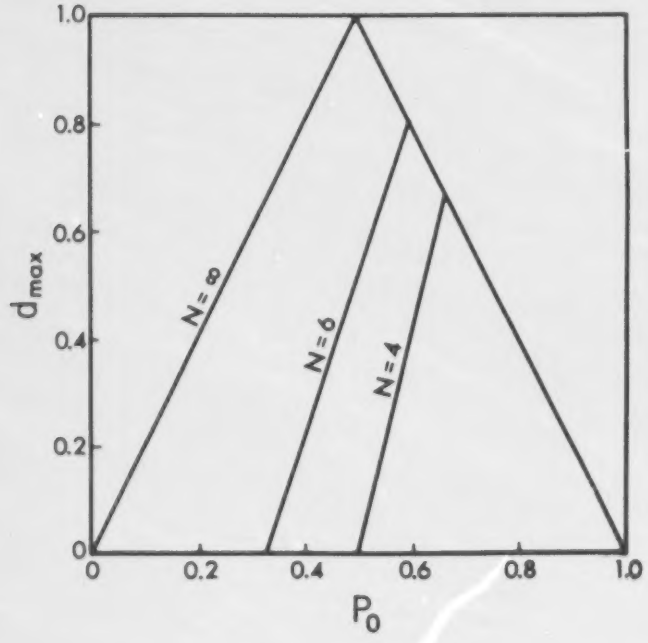
which is the simple guessing probability for N stimuli. (It is not $1/N$ because the right-ear response must be different from the left-ear response.) In other words, at this point a hypothetical listener with the maximal possible ear difference always can identify the stimulus in the dominant ear, but produces a random guess for the stimulus in the other ear. The maximal ear difference d_{max} at this point is

$$d_{max} = P_{Rmax} - P_{Lmin} = 1 - 1/(N - 1) = (N - 2)/(N - 1) \quad , \tag{15}$$

which is the maximal expected ear difference for a given N. It occurs at a performance level of



a.)



b.)

Figure 2: Maximal and minimal ear scores (upper panel) and their difference (lower panel) as a function of performance level, for three levels of guessing probability.

(16) $$P_o = 1/2 + 1/2(N - 1) = N/2(N - 1) .$$

From this point on, P_{Rmax} remains at 1.0 and P_{Lmin} increases with P_o. The complete functions relating P_{Rmax} and P_{Lmin} to P_o are:

(17) $$P_{Rmax} = [2/(N - 2)][(N - 1)P_o - 1] \quad \text{if } 2/N \leq P_o \leq N/2(N - 1)$$
$$P_{Rmax} = 1 \quad \text{if } N/2(N - 1) \leq P_o \leq 1.0 ,$$

(18) $$P_{Lmin} = [2/(N - 2)](1 - P_o) \quad \text{if } 2/N \leq P_o \leq N/2(N - 1)$$
$$P_{Lmin} = 2P_o - 1 \quad \text{if } N/2(N - 1) \leq P_o \leq 1.0 .$$

Figure 2b shows the corresponding relationship between d_{max} and P_o for N = ∞, N = 6, and N = 4. For a finite N, this function is

(19) $$d_{max} = P_{Rmax} - P_{Lmin} =$$
$$= [2/(N - 2)](NP_o - 2) \quad \text{if } 2/N \leq P_o \leq N/2(N - 1)$$
$$= 2(1 - P_o) \quad \text{if } N/2(N - 1) \leq P_o \leq 1.0 .$$

(The function for N = ∞ is given in Equation 3.)

Now we define e_g--as we will call e with the correction for guessing--as

(20) $$e_g = d/d_{max} =$$
$$= (P_R - P_L)/[2(NP_o - 2)/(N - 2)] \quad \text{if } 2/N \leq P_o \leq N/2(N - 1)$$
$$= (P_R - P_L)/[2(1 - P_o)] \quad \text{if } N/2(N - 1) \leq P_o \leq 1.0 .$$

Equation 20 shows that e_g is identical to e--and thus to POE'--in the upper range of performance levels. In other words, POE' is unaffected by guessing probability and needs no correction. It is only in the lower range of performance, where POC' applies, that a correction for guessing becomes necessary. Without it, the magnitudes of ear advantages at low performance levels would be seriously underestimated.

The correction for guessing that I have just proposed is only a global and approximate solution. Ideally, such a correction should be based on a detailed model of perceptual and response processes in dichotic listening. At present, such a model does not exist. Recently, I have considered a very simple probabilistic model that assumes that the listener either perceives a stimulus correctly or makes a random guess, independently for each ear. I found that the e index based on the resulting estimates of the "true" probabilities of perceiving left- and right-ear stimuli is almost identical to e_g. However, the model is too simple to provide a complete account of the perception of dichotic stimuli. A more detailed discussion of this approach is provided in a separate paper (Repp, 1977b).

1.5. Isolaterality Contours

The "correct" index of the ear advantage must fit both theoretical conceptions and empirical evidence. Halwes (1969) believed he had solved the theoretical problem by proposing an index (e) whose range is free of the constraints of performance level. However, this argument, intuitively appealing as it is, really attacked the problem from the wrong side, although it may have led to a correct outcome. Marshall et al. (1975), who also proposed e as

the perhaps best index, correctly stressed that different indices represent "psychological theories of how an S [subject] changes Rc and Lc [P_R and P_L] in achieving different overall accuracies" (p. 320). In other words, performance level must be understood as a consequence of changes in right-ear and left-ear scores, and the concomitant constraints on the ranges of certain indices must be accepted if they are predicted by theories about the form of covariation of P_R and P_L. There is no such theory that postulates that the range of an ear advantage index must not be constrained in any region of performance.

However, among the infinite number of possible theories, there is one class of theories that leads precisely to this outcome--an example is the theory underlying the e index. In order to clarify this point, consider the isolaterality contours assumed by different indices, that is, by different theories of the ear advantage. Isolaterality contours connect points of equal underlying ear asymmetry at different levels of performance. In Figure 1, these contours would be parallel horizontal lines within the limits of each index. It is more illuminating to represent these isolaterality contours in terms of P_R and P_L, as Marshall et al. (1975) have done. Figure 3 plots P_R against P_L, so that the isolaterality contours connect all pairs of scores (P_R, P_L) that are assumed to reflect the same underlying ear asymmetry. To simplify the exposition, we have assumed in Figure 3 that the guessing probability is zero; a nonzero guessing probability would have the effect of restricting the possible score combinations to a region in the upper right-hand corner of the unit square (or accuracy space, as Marshall et al. call it).

Figure 3 shows the isolaterality contours assumed by four theories: those associated with the indices d, POC', POE', and e. Note that the region above the positive diagonal represents REAs, while the symmetric region below the positive diagonal represents LEAs. The isolaterality contours are shown only for REAs; those for LEAs are obtained by symmetric reflection around the positive diagonal. The isoperformance contours, which connect all pairs of scores (P_R, P_L) at the same $P_o = (P_R + P_L)/2$, are straight lines parallel to the negative diagonal in each case.

Figure 3 shows that only the e index provides a definite estimate of the magnitude of the ear advantage for every pair of scores. The other three indices depicted can give only a lower or upper bound on the ear advantage when one of the two ear scores is either at chance level or perfect, because these data points cannot be uniquely assigned to a particular isolaterality contour. For example, the fact that d cannot exceed 0.2 when $P_o = .8$--due to the "constraint imposed by performance level on the range of the index," discussed in connection with Figure 1--really implies that, if d is the correct index (that is, if the theory underlying d is correct), any true ear advantage of $d > 0.2$ cannot be measured at $P_o = 0.8$. If the model underlying d happened to be correct, this disadvantage must be accepted; it cannot be taken as an a priori argument against the index-theory. Similar arguments apply to POC' and POE'.

From Figure 3, the close analogy to signal detection experiments that has also been pointed out by Marshall et al. (1975) is evident. P_L is formally analogous to the false alarm probability, and P_R to the hit probability in signal detection. Isolaterality contours correspond to receiver operating characteristic (ROC) functions, and isoperformance contours to isobias contours (cf. Green and Swets, 1966). The isolaterality contours assumed by the e

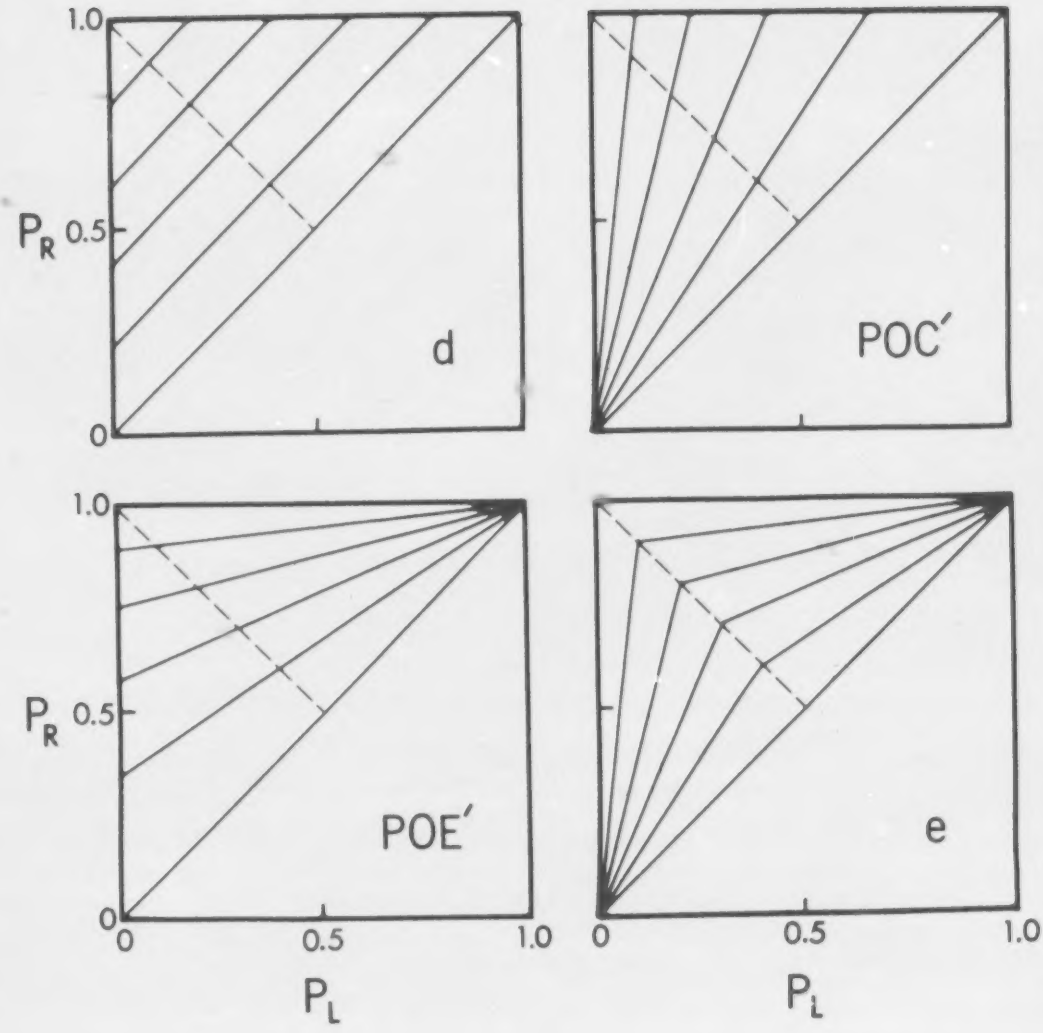


Figure 3: Isolaterality contours assumed by the theories underlying four indices of the ear advantage.

index (Figure 3d) are linear approximations to the ROC functions resulting from the standard signal detection model assuming underlying normal distributions with equal variance. This gives the e index some intuitive plausibility, in view of the success of the standard signal detection model in many different situations. However, whether it is also a correct model of dichotic listening remains to be proven. In the absence of stronger theoretical or empirical support, the alternative models underlying d, POC', or POE' cannot be ruled out. The POE' model, for example, corresponds to a "high-threshold" model in terms of signal detection theory that has been found useful in certain situations (Green and Swets, 1966). There is an infinity of other possible models; those depicted in Figure 3 are merely the extreme cases.

In addition to the intuitive appeal of e, its underlying assumptions may be plausibly conceptualized as follows. Assume that differences in performance level reflect different levels of noise in the perceptual-auditory system of listeners. Further, assume that, as the internal noise level is reduced from very high to very low, P_R and P_L increase independently of each other in the form of two ogive functions. The separation between these functions equals the true ear asymmetry and may be expressed in terms of the signal-detection statistic d'. This simple conception is identical with the standard signal detection model, so that e--whose isolaterality contours are a good approximation to the standard model--would be the correct index (if not d' itself is chosen as the index, which certainly is an option). Again, however, this argument has only intuitive plausibility at present. A decision between different models will require empirical evidence in favor of one or the other.

Unfortunately, empirical tests of the models are difficult. Marshall et al. (1975) have pointed out that, in analogy to signal detection, it would be necessary to vary performance level in a number of steps while holding the underlying ear asymmetry constant. This would generate points on the same ROC function whose shape could then be determined. There are both theoretical and practical problems with this approach. The most obvious technique would be to employ masking noise or some other form of distortion to vary performance level within a single subject, but it is not clear whether this variation would be equivalent to the hypothetical variations in internal noise level that cause variations in P_O between subjects, despite high monaural intelligibility of the stimuli. Cullen, Thompson, Hughes, Berlin, and Samson (1974) have varied signal-to-noise ratio in a dichotic two-response paradigm, but their results are irregular and permit no conclusion. A practical problem is that ear advantages tend to be rather small and highly variable, so that an enormous amount of data would be necessary to distinguish different shapes of ROC functions in the vicinity of the positive diagonal.

Halwes (1969) used the more global empirical approach of taking the average ear differences obtained for different groups of subjects in a number of different experiments and plotting them as a function of the "natural" variations in average performance level between the experiments. When the average ear advantages were expressed in terms of e, they turned out to be strikingly independent of performance level, which, at the time, provided impressive empirical support for the e index. Unfortunately, this result does not hold up in the light of more recent data. I have surveyed a large number of dichotic studies conducted since 1969 and found large variations in the magnitudes of ear advantages from study to study, regardless of performance

level, so that no clear conclusion emerges from the data.

Yet another way of testing the assumptions underlying the e index would be to conduct an analysis of individual stimulus pairs. A large amount of data would have to be collected for this purpose. If individual stimulus pairs vary in their "performance level," then the data points (P_R plotted against P_L) for all individual stimulus pairs should lie on the same ROC function. This approach is analogous to that discussed in Section 3 for the single-response paradigm, and it is certainly worth investigating. However, it is not clear whether individual stimulus pairs vary more than randomly in "performance level" (except for the "feature-sharing effect" discussed in Section 4); performance level has so far been considered a characteristic of the listener, not of the stimuli. More detailed investigations of the dichotic competition between individual stimuli are needed.

Thus, although e has the advantage of being the most intuitively satisfying index, other indices and their corresponding models cannot be ruled out completely at present. I would recommend, however, that e_g (that is, e with the correction for guessing--Equation 20) be adopted as an index as long as there is no evidence that speaks against its use. In the remainder of this paper, we will describe a simpler approach to measuring the ear advantage that, despite many analogies, avoids some of the problems inherent in the two-response paradigm. Some of these problems will become clear as the discussion proceeds (see especially Section 4). Considering the complexity of the two-response paradigm, it may be time to look for alternative methods that perhaps achieve the same goal with fewer complications.

2. DICHOTIC FUSION AND THE SINGLE-RESPONSE PARADIGM

2.1. Dichotic Fusion

When two sounds are presented simultaneously to the two ears, they are not always perceived as two separate events. Often they fuse into a single sound image. This is obviously true when the two sounds are exactly identical. In real life, environmental sounds normally reach both ears, but the signals at each ear typically show slight differences in spectrum, intensity, and time of onset. Nevertheless, they give rise to a single localized sound image (Mills, 1972).

Stereo headphones make it possible to present different sounds independently to the two ears and thus to investigate the mechanisms of binaural (dichotic) fusion. Laboratory studies have shown that the fusion mechanism tolerates a certain amount of spectral discrepancy beyond that encountered in natural situations. For example, dichotic sinusoids within a certain critical frequency range (the "binaural critical band") are heard as a single tone, although it may "beat" when low frequencies are involved (Odenthal, 1963; Perrott and Barry, 1969; Van den Brink, Sintnicolaas, and Van Stam, 1976). The width of the binaural critical band increases with signal frequency (Perrott and Barry, 1969) and intensity (Perrott, 1970); it also increases as the signal duration decreases (Perrott, Briggs, and Perrott, 1970). The fused tone is heard at a frequency intermediate between the two dichotic frequencies (Odenthal, 1963). Of special importance is the finding that two different tones that normally would not fuse can be made to fuse by imposing the same low-

frequency modulation onto them (Leakey, Sayers, and Cherry, 1958; Tobias, 1972). In general, it seems that complex auditory signals with similar waveform envelopes fuse despite considerable differences in microstructure.

This result is important in the dichotic fusion of speech sounds. The waveform envelope of a speech signal is determined by its low-frequency components (primarily the fundamental frequency), while the higher formants constitute the microstructure. Two different formants presented dichotically at the same fundamental frequency fuse into a single sound, while two formants with the same center frequency but with different fundamental frequencies are heard as separate sounds (Broadbent and Ladefoged, 1957). Thus, a speech signal may be "split" by filtering it into nonoverlapping low- and high-frequency bands which, if presented simultaneously to the two ears, are heard as a single source resembling the original (Broadbent, 1955; Franklin, 1969). Several recent studies have employed the related "split-formant technique" with synthetic speech, where some formants are presented to one ear and the remaining formants to the other ear (Rand, 1974; Nye, Nearey, and Rand, 1974; Nearey and Levitt, 1974; Haggard, 1975). Cutting (1976), in his recent classification of dichotic fusion phenomena, called this "spectral fusion."

Dichotic fusion is not limited to the case where parts of a speech signal fuse to reconstitute the original whole stimulus. Even if two different complete utterances are presented, the perceptual result may be a single fused stimulus, provided that the two dichotic stimuli have sufficiently similar fundamental frequencies. The fused percept may resemble one or the other component, or it may be a hybrid (see Cutting, 1976). In assessing dichotic ear differences, it is important to know whether some or all of the stimuli fuse. Ideally, the experimenter should be able to control this property of the stimuli.

The verbal materials used in dichotic listening studies may be roughly classified into three groups:

(1) Words, digits, and other larger-sized verbal units. Typically, they are natural speech and acoustically heterogeneous, so that the waveforms in the two ears show little correspondence. Therefore, they tend not to fuse.[2]

(2) Natural-speech nonsense syllables that have been used extensively in recent research (for example, Studdert-Kennedy and Shankweiler, 1970; Berlin, Lowe-Bell, Cullen, Thompson, and Loovis, 1973; Cullen, Thompson, Hughes, Berlin, and Samson, 1974). The typical set is /ba/, /da/, /ga/, /pa/, /ta/, /ka/, spoken by the same voice. Some of the dichotic pairs formed from these syllables may fuse into a single syllable if they are spectrally similar and properly synchronized; this will depend on the particular stimuli and recording procedures used. Apart from temporal alignment, however, the experimenter has little control over fusion. Tests of this kind often contain fused and unfused

---

[2]Nevertheless, the spectral separation of the two competing signals may affect performance. Perceptual separability may be viewed as a continuum ranging from perfect fusion to perfect separability. (See also the discussion of selective attention in Section 4.4)

pairs mixed together, which is a methodological disadvantage.

(3) Synthetic syllables (for example, Halwes, 1969; Shankweiler and Studdert-Kennedy, 1967, 1975). As with any other stimuli, it depends on their spectral similarity (most of all on their fundamental frequencies) whether they do or do not fuse. However, the important advantage of synthetic syllables is that their acoustic properties--and, hence, their tendency to fuse--are under the control of the experimenter. Thus it is possible to construct homogeneous tests that contain only pairs that fuse, or only pairs that do not fuse.

The most widely used synthetic stimulus set is /ba/, /da/, /ga/, /pa/, /ta/, /ka/ with identical fundamental frequency contours. As with the analogous natural-speech set of syllables, the reason for their popularity is primarily the convenience and availability of a stimulus set that tends to give reliable REAs--not their tendency to fuse, that has been given little attention. The differences between these stimuli are confined to the first 50 msec or so, which carry the consonantal distinctions. The vowel portions--that may last for another 250 msec or so--are exactly identical and therefore fuse perfectly in dichotic presentation. This alone is sufficient to guarantee that dichotic pairs of these stimuli will sound more or less fused (Halwes, 1969). The "more or less" will depend on the spectral similarity of the initial 50 msec. Synthetic /ba/, /da/, /ga/, if synthesized so they differ only in the transitions of the second (and third) formant fuse perfectly into a single syllable. This was experimentally demonstrated by requiring subjects to discriminate dichotic pairs from binaural (identical) pairs of stimuli from the same set. Most of the subjects, including experienced listeners, performed at chance level (Repp, 1976b). It is justified, therefore, to call these stimulus pairs "perfectly fused".[3]

Informal observations suggest that strong fusion is also obtained for the voiceless set (/pa/, /ta/, /ka/) if the stimuli differ only in their formant transitions. On the other hand, stimuli that contrast in voicing (and thus in the relevant cue, voice onset time, so that a periodic waveform in one ear is accompanied by filtered noise in the other ear during the first 50 msec or so) are sufficiently different to prevent perfect fusion. The listener has some indication that different events have occurred in the two ears, but since these events are immediately followed by a perfectly fused vowel, their discrepancy is perceived only as a brief noise or roughness accompanying the perception of a single fused syllable that can be identified without great difficulty. Dichotic pairs consisting of a single phonetic percept accompanied by an auditory signal of interaural discrepancy may be called "partially fused".

The fusion of synthetic syllables can be effectively prevented by presenting them at different fundamental frequencies (Halwes, 1969; Repp, 1976a). Temporal asynchrony also reduces fusion, but as long as the signals overlap, they may still partially fuse. Some researchers have paired CV syllables that contrasted in their vowels as well as in the initial consonants (Studdert-

---

[3]The actual syllables in this experiment were /bae/, /dae/, /gae/, but the nature of the vowel is immaterial. Perfectly fused syllables have been used in a number of other studies since (Repp, 1976c, and unpublished work).

Kennedy, Shankweiler, and Pisoni, 1972). Different vowels with the same fundamental frequency seem to fuse quite well, although they may be discriminable from binaural stimuli if they are spectrally dissimilar (Kuwahara and Sakai, 1976).[4] The frequency of the first formant may play a role in addition to fundamental frequency, but little work has been done on the fusion of complex sounds such as vowels. The influence of various other parameters, such as differences in initial bursts, transition duration, etc., on dichotic fusion of speech sounds has not been systematically studied. If stimuli involving such differences are to be used for assessing ear advantages, their degree of fusion should first be determined.

2.2. The Single-Response Paradigm

The standard procedure requires the subjects in a dichotic test to identify both competing stimuli. While appropriate with unfused stimuli, the two-response procedure has also been used with synthetic syllables subject to dichotic fusion (for example, Shankweiler and Studdert-Kennedy, 1975). It is not surprising that the overall accuracy was quite low in these studies, because at least one of the two responses must have been a guess. Although it is possible to analyze only first responses and ignore second responses, one cannot be sure that the subjects always record their most confident response first, even when instructed to do so. Thus, the responses reflecting what the listeners actually perceived are distributed over two response columns, and it is impossible for the experimenter to identify them reliably. Hence, instructions to identify two stimuli when only one is heard are inappropriate. The only appropriate instruction is simply to identify the syllable heard (the single-response paradigm). The listener need not even be informed about the presence of different events in the two ears. Instructions to selectively attend to one ear are also inappropriate when the stimuli are fused, since it has been shown that selective attention to one ear has little or no effect with fused stimuli (Halwes, 1969; Repp, 1976b). The topic of selective attention will be discussed in more detail in Section 4.4.

Thus, dichotic tests using fused syllables are quite different from those using unfused stimuli. With unfused stimuli, the subject gives two responses that are then classified as correct or incorrect. The emphasis is on accuracy of identification. A large number of errors is desirable. These errors should be due to dichotic competition only; the monaural intelligibility of the stimuli should be as high as possible. The "raw" ear advantage (d) is defined as the difference between the proportions of correct responses for the two ears.

In a test using fused stimuli, on the other hand, only a single response is given to each stimulus pair. Ideally, this response should match one or the other of the component stimuli. Dichotic pairs for which this indeed tends to be the case (for example, /ba/-/da/, which is heard as either /ba/ or /da/) are especially desirable. Other pairs also yield hybrid responses such as "psychoacoustic fusions" or blend responses (Cutting, 1976; Repp, 1976b, 1977a). The

---

[4]Kuwahara, H., and Sakai, H. Identification and dichotic fusion of time-varying synthetic vowels. Unpublished manuscript, 1976.

methodological problems created by such responses will be discussed in Section 4. If we consider only the "ideal" pairs, such as /ba/-/da/, where virtually all responses match one of the two component stimuli, we see that there are no errors and accuracy is perfect (or, in practice, as good as the monaural intelligibility of the stimuli). The question is not how accurately each ear performed, but how the competing information was weighted and combined into a single perceptual outcome. Thus, the emphasis is on dichotic integration, not on competition. Instead of different accuracy levels for each ear, we have two complementary proportions representing each ear's share of the responses. The difference between these proportions represents the "raw" ear advantage.

Despite the theoretical and methodological differences, the two paradigms also have much in common. Specifically, the problems encountered in deriving an appropriate laterality index are rather similar. This will become evident in the following section which derives such an index for the single-response paradigm.

3. LATERALITY INDICES FOR THE SINGLE-RESPONSE PARADIGM

3.1. Ear Dominance and Stimulus Dominance

In this section, we make the simplifying assumption that each stimulus pair in a test using fused stimuli yields only two kinds of (single) responses, one that matches the stimulus presented to the left ear, and one that matches the stimulus in the right ear. One example, already mentioned in the preceding section, is the pair /ba/-/da/, which is heard as either /ba/ or /da/. (For other examples, see Section 4.5.) Thus, the responses can be divided into those reflecting perceptual dominance of the left-ear stimulus and those reflecting perceptual dominance of the right-ear stimulus. Taking into account the two possible channel/ear assignments of the stimuli, the data for a single stimulus pair can then be represented in a 2 x 2 table, as illustrated in Table 2. The two different channel/ear assignments of the stimuli constitute the rows of this table, and the two responses the columns. The entries are the proportions of the two responses for each of the two channel configurations.

TABLE 2: The data structure for a single stimulus pair in the single-response paradigm, with sample values.

Channels		Responses		
LE	RE	/ba/	/da/	
/ba/	/da/	$x_i = 0.276$	$1 - x_i = 0.724$	1.000
/da/	/ba/	$y_i = 0.487$	$1 - y_i = 0.513$	1.000

Perceptual dominance is a probabilistic phenomenon, so that, in general, both responses will occur with some frequency over a number of single-response trials. There are two independent factors that determine which of the two competing stimuli dominates the perception of the fused syllable at a given

time. One is the tendency of (the stimulus in) one ear to dominate (the stimulus in) the other ear. It is appropriately called ear dominance and, of course, is analogous to the ear advantage observed in the two-response paradigm.[5] The other factor is the tendency of one stimulus to dominate the other stimulus, regardless of their particular channel assignment. It may be called stimulus dominance and constitutes an important phenomenon in its own right (Repp, 1976b).

The two factors are illustrated by the fictitious data in Table 2. Ear dominance is reflected in the difference between the averages of the diagonal entries in the 2 x 2 table. In the present example, there is a right-ear dominance: (72.4 + 48.7)/2 = 60.5 percent of the responses went to the right ear and only (27.6 + 51.3)/2 = 39.5 percent to the left ear. At the same time, there is a pronounced stimulus dominance effect, which is reflected in the difference between the column averages: /da/ was heard in (72.4 + 51.3)/2 = 61.8 percent of the trials; /ba/ only in (27.6 + 48.7)/2 = 38.2 percent.[6]

It should be emphasized that the information about ear and stimulus dominance is contained only in the complete 2 x 2 contingency table but not in its individual rows. The two different channel assignments of a particular stimulus pair must always be considered together; otherwise, the results can be very misleading. In Table 2, for example, /da/-/ba/ (with /ba/ in the right ear) shows a slight LEA, while /ba/-/da/ (with /da/ in the right ear) shows a very large REA. Such a result can appear puzzling if it is interpreted without awareness of the joint operation of two factors, ear dominance and stimulus dominance (cf. Speaks, Niccum, Carney, and Marble, 1975; Niccum, Speaks, and Carney, 1976). In fact, the right-ear dominance underlying these data is cancelled by stimulus dominance in the pair /da/-/ba/, and it is augmented by stimulus dominance in the pair /ba/-/da/. Neither case in isolation reveals the actual size of the REA which lies between these extremes and must be inferred from the complete contingency table. Likewise, an appropriate estimate of stimulus dominance in an individual stimulus combination can only be derived from the complete table.

---

[5] In order to avoid new acronyms, the abbreviations REA and LEA will be maintained for the corresponding trends in ear dominance.

[6] It may be argued that stimulus dominance reflects merely response bias, that is, a stimulus-independent tendency of listeners to give one response more often than the other. However, stimulus dominance relationships can be changed by modifying the acoustic structure of the stimuli within phonetic categories (Repp, 1976b, 1977b), so that they are at least in part stimulus-dependent. Repp (1976b) hypothesized that stimulus dominance is completely determined by the relationship of the stimuli to the listener's perceptual category prototypes. Essentially, this is a theory of response bias. Stimulus dominance may be considered as the result of the interaction between the listener's perceptual organization and the structure of the stimuli.

Table 2 bears a close resemblance to the 2 x 2 contingency table for the two-response paradigm (Table 1). However, in Table 1, the dimensions were left/right ear and correct/incorrect responses. The analogy becomes closer if one response in Table 2 is arbitrarily considered as "correct" (for example, /ba/) and the other as "incorrect" (for example, /da/). The rows of the two tables remain incompatible; however, in Table 1, they represent the individual component stimuli in each ear, while in Table 2 they represent the two possible channel/ear assignments of both component stimuli. In the two-response paradigm, it is easy to summarize the responses to all stimulus pairs in a single table; in fact, it is standard procedure to do so, and the data are rarely broken down to the level of individual stimulus pairs. Basically, each individual channel assignment of each stimulus pair yields its own 2 x 2 table (of the form shown in Table 1), and these tables are then simply added up or averaged. This presents no problem, because each stimulus pair yields left-ear and right-ear as well as correct and incorrect responses. In the single-response paradigm, on the other hand, the 2 x 2 tables for the individual stimulus pairs are not commensurate--their rows and columns have different labels in each case--and therefore cannot simply be added up or averaged. Even if we stipulate that the positive diagonal always contain right-ear responses and the negative diagonal left-ear responses (as in Table 2), there remains one degree of freedom for the arrangement of the table. We show now how this problem can be solved.

3.2. The e Index for the Single-Response Paradigm

The problem now at hand is how to compute an appropriate laterality index for a whole single-response test. It is easy to compute ear dominance indices for individual stimulus pairs. Despite the different nature of the entries in Table 1 and Table 2, the structure of the data is almost completely identical in the two cases, and most of the discussion of Section 1 applies. In particular, the factor of stimulus dominance exacts the same constraints here as the factor of performance level in the two-response paradigm. A 50/50 distribution of responses here is analogous to a 50 percent performance level there. The simple difference index, $d_i = y_i - x_i$, is unsatisfactory for the same reason that d is unsatisfactory in the two-response paradigm. (The subscript i indicates that we are dealing with a single stimulus pair.) Clearly, the best choice is

$$(21) \qquad \begin{aligned} e_i &= (y_i - x_i)/(y_i + x_i) && \text{if } (y_i + x_i)/2 \leq 0.5 \\ &= (y_i - x_i)/(2 - y_i - x_i) && \text{if } (y_i + x_i)/2 \geq 0.5 \quad . \end{aligned}$$

Since the arrangement of the 2 x 2 data table is arbitrary, the convention of tabulating the less frequent response in the left column (as in Table 2) may be adopted, so that the first condition always holds and

$$(22) \qquad e_i = (y_i - x_i)/(y_i + x_i) \quad .$$

Thus, a laterality index can be computed for each individual stimulus pair. The most straightforward way of arriving at an index for the whole test would then be to take the average of all the e_i indices. However, these indices vary considerably in their precision, depending on how much stimulus dominance deviates from equilibrium. The e_i indices are most reliable when the two stimuli are in equilibrium, and they become more variable and unreliable as the relative dominance of one or the other stimulus increases.

This follows straightforwardly from statistical arguments. Therefore, e_i indices for stimulus pairs with very asymmetrical response distributions should receive less weight than indices for stimulus pairs with more nearly symmetrical response distributions. The degree of asymmetry is represented by the proportion of the less frequent of the two responses, $w_i = (y_i + x_i)/2$, which is the appropriate weight to be assigned to each e_i. The overall E index as the weighted average of the e_i indices is then computed,

$$(23) \qquad E = \sum w_i e_i / \sum w_i = \\ = (1/2) \sum [(y_i + x_i)(y_i - x_i)/(y_i + x_i)] / (1/2) \sum (y_i + x_i) = \\ = \sum (y_i - x_i) / \sum (y_i + x_i) = e \ .$$

Thus, the result turns out to be identical with the e index computed from a summary 2 x 2 table for the whole test. We note that, by adopting the conventions of tabulating the less frequent response in the left column and right-ear responses in the positive diagonal, we have fixed the format of the data tables, so that they can now be added up or averaged in a nonarbitrary way. The e index computed from this summary table is then identical with the weighted average of the e_i indices for the individual stimulus pairs.

The variance of the e_i indices provides us with an estimate of whether the overall e index is significantly different from zero. Assuming that the e_i indices are approximately normally distributed around zero if the null hypothesis is true, we make use of the well-known relation that the estimated variance of the mean is the sample variance divided by the number of observations,

$$(24) \qquad s^2(e) = s^2_w(e_i)/N \ ,$$

where N is the number of stimulus pairs. The subscript w indicates that, again, we would like to assign more weight to the deviations of the more reliable indices from the mean than to the deviations of unreliable indices. We thus compute the weighted variance of the e_i indices as

$$(25) \qquad s^2_w(e_i) = \sum w_i (e_i - e)^2 / \sum w_i = \\ = \sum w_i e_i^2 / \sum w_i - e^2 = \\ = \sum [(y_i - x_i)^2 / (y_i + x_i)] / \sum (y_i + x_i) - \\ - [\sum (y_i - x_i) / \sum (y_i + x_i)]^2 \ .$$

Confidence limits for e can then be estimated by $e \pm 2s(e)$. If they do not include zero, e is significant at approximately $p < .05$.

3.3. The e' Index

The e index will be useful as long as the distribution of the e_i indices is roughly symmetrical. With very asymmetric distributions, however, an arithmetic mean is not the optimal measure. There is an alternative method available that also permits an approximate graphical determination of the laterality index. This method uses the basic concepts of signal detection theory that have already been referred to in Section 1. In addition, it provides a direct test of the assumptions underlying the e index. The procedure is illustrated in Figure 4 using some actual data from a recent experiment by Repp (1977a).

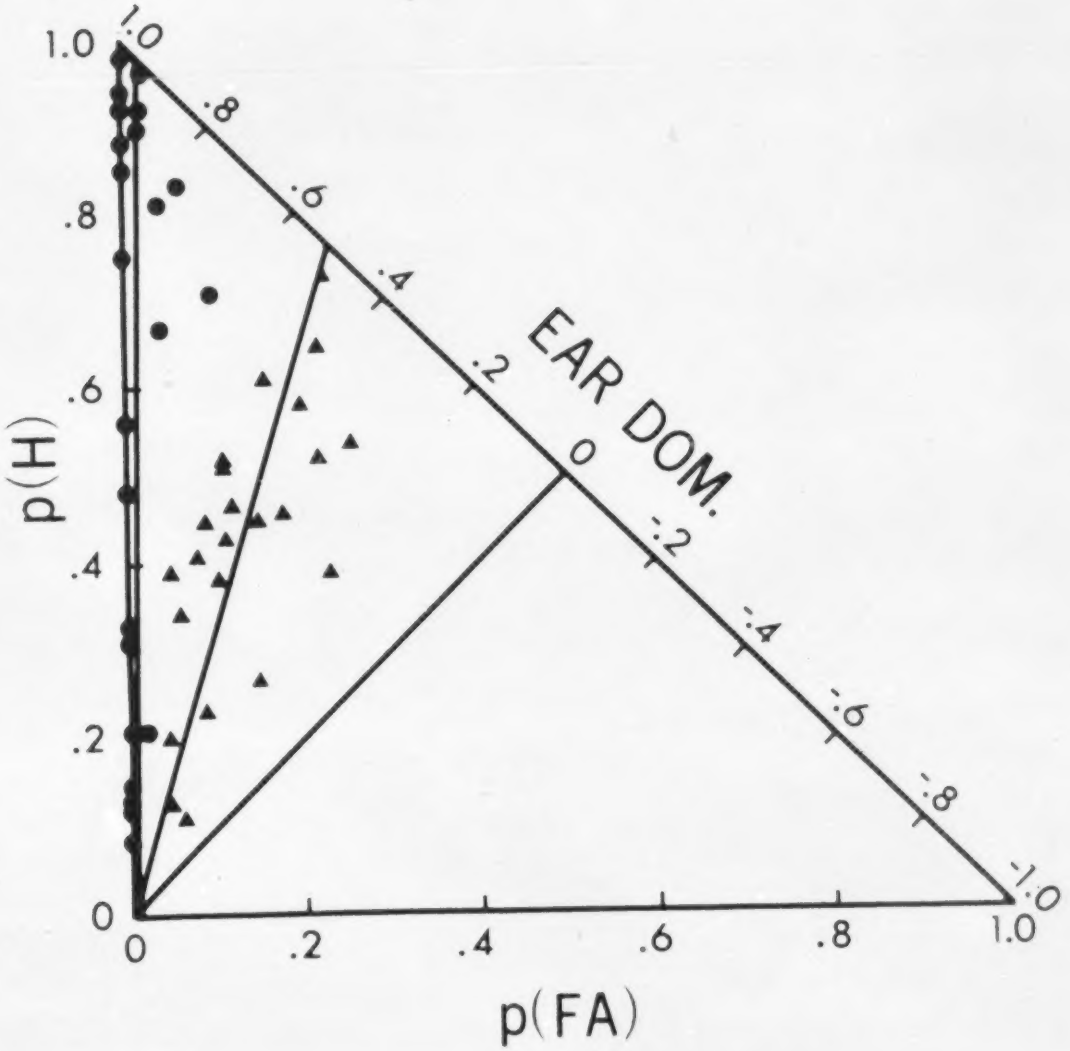


Figure 4: Illustration of the graphic derivation of the e' index. Data from Repp (1977a). See text for explanation.

Attention is again restricted to the less frequent responses only, that is, to the left columns of the N data tables for the individual stimulus combinations. The entry in the top row, x_i, represents left-ear responses, or "false alarms." The entry in the bottom row, y_i, represents right-ear responses, or "hits." We then plot y_i--or p(H), the hit probability--against x_i--or p(FA), the false alarm probability--for all stimulus pairs. This results in a swarm of points located on or below the negative diagonal of the unit square. (Therefore, only its lower triangular portion is shown in Figure 4.) To these points, a receiver operating characteristic (ROC) function may be fitted. The standard ROC function is curvilinear, but for our purposes little accuracy is lost by simply fitting a linear function. A straight line through the origin and the data points may be fitted by eye, or, more precisely, by the method of least squares. The slope b of this line will range from infinity (perfect REA) to 0 (perfect LEA). In order to convert this range to the standard scale from +1 to -1, we define

$$e' = (b - 1)/(b + 1) . \tag{26}$$

This value can also be read off a linear scale on the negative diagonal, as illustrated in Figure 4. The triangles are the average results of eight subjects, while the circles are for a single experienced listener (myself) who showed an especially large REA. Based on 24 data points (stimulus pairs) in each case, the e' coefficients are +0.55 and +0.96, respectively.

e' may be directly calculated as

$$e' = \tan[(1/2)\arctan[(\sum y_i^2 - \sum x_i^2)/2 \sum x_i y_i]] , \tag{27}$$

which effectively is a rotation of the best-fitting line into the ± 45 degrees sector, so that its slope (the tangent) ranges from +1 to -1.

The e' index is an unbiased measure in terms of signal detection theory, since it is a simple linear transformation of the area under the ROC function, a commonly used measure of sensitivity that is independent of any particular assumptions about the internal representations of the sensory events (Green and Swets, 1966; Richardson, 1972). Testing the significance of e' is not straightforward, so that one may rely on the e approximation (Equation 22) for this purpose.

TABLE 3: Ear advantages on the voicing dimension. Data of eight subjects from Repp (1977a).

Subjects	e'	e	s(e)
JK	+0.17	+0.16	0.06
JL	+0.73	+0.72	0.06
RG	+0.89	+0.89	0.02
MR	+0.57	+0.49	0.10
GG	-0.09	-0.12	0.08
WT	+0.90	+0.88	0.04
TJ	+0.47	+0.44	0.07
CW	+0.75	+0.74	0.06

The e' index is usually well approximated by e. Table 3 presents e' and e coefficients, together with s(e), for the eight subjects in Repp's (1977a) study. It can be seen that e is generally very close to e'; the largest deviation occurs for subject MR, who, in fact, showed a highly asymmetrical distribution of e_i indices. By the $\pm 2s$ criterion, all coefficients except that for subject GG (the only case of left-ear dominance) are significant. It should be noted that s(e) becomes constrained as e approaches ± 1 (cf. subjects RG and WT in Table 3), so that it should not be used for testing whether two coefficients are significantly different from each other. A nonparametric test may be used for this purpose.

One important difference between the present procedure of deriving e' and the signal detection paradigm should be pointed out. In the latter, "bias" is varied by means of instructions, payoffs, etc., while the stimuli for which sensitivity is being measured are held constant. If the stimuli (for example, signal and/or noise levels) were to be changed, the listener's sensitivity would change, too. In the present case, stimulus dominance takes the role of bias, and ear dominance that of sensitivity. However, in order to change stimulus dominance, the stimuli themselves are varied. Thus, it is assumed that ear dominance is independent of the nature of the stimuli, at least within a given class (such as initial stop consonants). The validity of this assumption is an empirical question. It is especially convenient that determining e' for a set of data at the same time provides a test of its underlying assumptions: if the linear ROC function fits the data poorly, a different function and a different index may have to be chosen. So far the results have been encouraging.[6a] Moreover, no correction for guessing is needed for e' since, in general, guessing plays only a small role in the single-response paradigm. However, the single-response paradigm is not without its own problems. The last section discusses a number of methodological issues and problems so far not considered.

4. PROBLEMS IN MEASURING THE DICHOTIC EAR ADVANTAGE

4.1. Stimulus Intelligibility

It is good practice to precede a dichotic test with a series of binaural (or monaural) stimuli, in order to familiarize the listener with their sound and to find out whether they can be reliably identified. In order to obtain useful dichotic data, the stimuli must be intelligible and yield high binaural (or monaural) identification scores.

This goal is more easily achieved with natural speech stimuli than with synthetic speech. However, synthetic stimuli are desirable because their acoustic properties can be controlled by the experimenter. Therefore, it is advisable to use a good set of synthetic stimuli that has been pretested for intelligibility--a point that has often been neglected in the past.

Even when the average intelligibility of a set of syllables is high, their intelligibility should be tested for each individual subject in a given

---

[6a] Repp, B. H. Stimulus dominance and ear dominance in the perception of dichotic voicing contrasts. (submitted for publication).

test. From time to time, individuals are encountered who find it very difficult to identify synthetic speech sounds. Such individuals may have to be excused from the test. (This is an obvious problem in clinical applications of dichotic tests.)

Intelligibility is usually assessed in terms of the confusions that occur between members of a stimulus set. The information obtained from a monaural confusion matrix may be used to apply a correction to dichotic data that leads to a better estimate of the stimulus dominance relationships between the stimuli (Repp, 1976b). Unfortunately, however, information about ear dominance cannot be recovered in this fashion--confusable stimuli yield smaller ear advantages than nonconfusable stimuli (Repp, 1977a). Since this effect may be confounded with individual differences in confusion patterns, it is advisable to omit confusable pairs when calculating ear advantage indices for individual subjects.

Problems arising from confusability of certain stimuli may also be reduced by using a dichotic listening procedure that does not require a labeling response. Only one such alternative is mentioned here, originally proposed by Preston, Yeni-Komshian, and Benson (1968), and especially suited for fused stimuli: the two component stimuli are presented binaurally, followed by the dichotic pair, and the listener judges whether the dichotic stimulus was more similar to the first or the second binaural stimulus. I am currently experimenting with an AXB version of this ABX paradigm, that is, with the dichotic pair in the middle of each stimulus triad (cf. Repp, 1976a). This method may yield cleaner data than the single-response identification task, but it is more time-consuming.

The intelligibility-confusability issue raises an important theoretical problem. Individual differences in the perception of stimuli (especially of synthetic syllables) are large, and "poor subjects" who produce many confusions will tend to have smaller ear advantages than "good subjects". The individual differences that thus confound the measure of the ear advantage may be ascribed to different levels of "internal noise" in the listeners' perceptual systems. Now suppose we have succeeded in generating an excellent stimulus set that produces no confusions at all. Have we eliminated the individual differences? Overtly, yes; but if the stimuli were attenuated or mixed with white noise, some subjects probably would produce more confusions than others. Also, if tested with an acoustic stimulus continuum as used in categorical-perception studies, some subjects would have sharper category boundaries than others. Again, this may be ascribed to individual differences in internal noise level--most most likely the same differences that are evident with confusable stimuli.

Given such individual differences in perceptual accuracy, it is likely that they play a role in the perception of dichotic stimuli. A fused dichotic syllable pair is often quite ambiguous, and an unfused stimulus pair is often degraded through mutual interference between the two stimuli. The problem is best illustrated with a fused pair, for example, /da/-/ga/, as shown in Figure 5. Assuming perfect intelligibilty of the component stimuli and no pronounced stimulus dominance effect, this dichotic pair sometimes sounds like /da/ and sometimes like /ga/. (Because of categorical perception, the subject may often not be aware of the inherent ambiguity of the syllable.) For individuals with a REA, the dichotic pair sounds a little more

(often) like /da/ when /da/ is in the right ear, and a little more (often) like /ga/ when /ga/ is in the right ear. Thus, the two fused stimuli may be considered as lying on a /da/-/ga/ continuum, a little to the left and a little to the right of the category boundary, respectively. A "good subject" with a low internal noise level has a sharp category boundary and thus resolves the two dichotic stimuli well; he or she will show a clear REA. A "poor subject", on the other hand, with the same underlying REA as the good subject, is likely to have a flatter psychometric function separating the two categories and, as a result, will produce similar response distributions for the two dichotic pairs and a much smaller REA. This is schematically illustrated in Figure 5.

If this argument is correct, it implies that individual differences in the dichotic ear advantage may be inextricably confounded with individual differences in internal noise level. This would be a serious obstacle to measuring individual ear advantages on an ordinal scale.

This problem seems to be less acute in the two-response paradigm; there, variations in perceptual accuracy are translated primarily into variations in performance level that can be dealt with more easily. However, this apparent advantage of the two-response paradigm is offset by a number of disadvantages that are discussed in the next paragraphs.

4.2. Stimulus Dominance

The factor of stimulus dominance can be dealt with elegantly in the single-response paradigm, thanks to the analogy with performance level in the two-response paradigm. However, this analogy is purely formal--these are the factors that can be effectively handled in the respective paradigms by using similar methods--but they are conceptually very different. As the preceding paragraphs have shown, presumably there are variations in "performance level" (that is, perceptual accuracy) in the single-response paradigm, but they are covert and much more difficult to deal with. Correspondingly, there is the problem of how to deal with stimulus dominance in the two-response paradigm.

Although stimulus dominance may be expected to play a smaller role in the two-response paradigm, there is evidence that it is nevertheless present. Berlin et al. (1973), for example, have reported that unfused natural-speech syllable pairs that contrast in voicing receive more correct voiceless responses than correct voiced responses. Berlin et al. reduced this asymmetry by aligning the stimuli at the first pitch pulse rather than at stimulus onset. Speaks et al. (1975), who used the same alignment criterion, reported data suggesting that stimulus dominance effects are of minor importance with natural-speech stimuli. However, this issue needs to be investigated in more detail. There is no doubt that strong stimulus dominance reduces the manifest ear advantage, and if some individuals show stronger effects than others, these individual differences are confounded with the measure of the ear advantage. At present, I know no way of dealing with this potential problem.

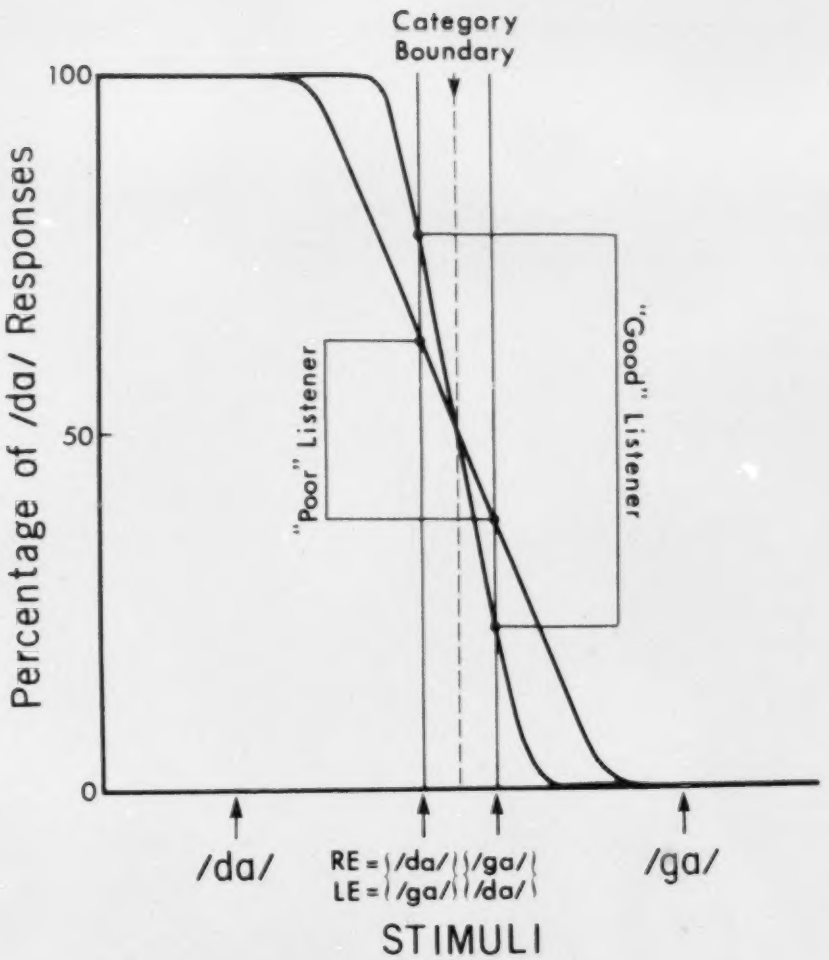


Figure 5: Schematic illustration of the effect that individual differences in perceptual acuity might have on ear dominance for fused syllables.

4.3. Guessing

Guessing plays an insignificant role in the single-response paradigm. Random guesses following lapses of attention or highly ambiguous stimuli may occur now and then, but, in general, the listener reports what he or she hears and does not resort to guessing (except for "sophisticated" guessing between a very limited number of alternatives). In the two-response paradigm, on the other hand, guessing is commonplace. Frequently a listener can identify the stimulus in one ear but has no clues about the stimulus in the other ear. The resulting guesses cannot be reliably identified in the data and lead to a considerable amount of random variation. The correction for guessing proposed in Section 1.2 is a rather crude procedure, and alternative ways of dealing with the guessing problem should be considered.

One obvious possibility is to instruct the listeners not to guess, that is, to give zero, one, or two responses per stimulus pair, depending on how many stimuli he or she heard clearly. This method has rarely been used, because the resulting heterogeneous protocols are difficult to analyze. More common instructions have been to write down the more confident response first and to analyze only these responses. Effectively, this is the single-response paradigm applied to unfused stimuli. (The second response might just as well be omitted.) If both stimuli can be identified, it amounts to a judgment which of them is the more salient. This procedure is interesting because it reduces guessing and permits the methods of Section 3 to be applied so that stimulus dominance can be taken into account. The main problem is the control of selective attention, discussed in the next paragraphs.

4.4. Selective Attention

The most important difference between fused and unfused syllables lies in the effect of selective attention. Perfectly fused syllables are heard as originating in the middle of the head, and voluntary efforts to pay attention to one ear has no effect on the responses (Repp, 1976b). The effect of selective attention with partially fused syllables appears to be very small, although this issue deserves further investigation (see Halwes, 1969; Repp, 1976a). Unfused syllables, on the other hand, yield large attentional effects, and practiced listeners are able to reach almost perfect scores when reporting only the syllables in one ear (Halwes, 1969). It is fair to say that the effectiveness of selective attention is a direct function of the degree of fusion of two stimuli (cf. also Footnote 2).

It follows that, with unfused syllables, it is not possible to separate attentional preferences, effectiveness, or bias from the ear advantage per se that presumably has a physiological basis. Some researchers have hypothesized that the ear advantage is entirely an attentional phenomenon (Kinsbourne, 1973, 1975; Morais and Landercy, 1977) or a perceptual advantage for stimuli localized to the right of the midline (Morais and Bertelson, 1973, 1975; Morais, 1975; Hublet, Morais, and Bertelson, 1976). The fact that a REA is obtained for perfectly fused syllables in the absence of any attentional effects (Repp, 1976b, 1976c) suggests that there are both physiological and attentional components of the ear advantage. Perfectly fused syllables may yield an estimate of the physiological component alone, with the attentional component removed. This makes tests using fused

syllables such promising instruments. With unfused syllables, physiological and attentional effects are confounded.

In fairness, one should distinguish two kinds of attentional effects: automatic and strategic biases. Automatic biases may arise from contingencies and expectancies within the experimental situation; for example, during or after processing a verbal stimulus, the left hemisphere may be activated more than the right, leading to an automatic bias for stimuli on the right side. These kinds of involuntary biases are what Kinsbourne and Morais have in mind. The REA for unfused syllables apparently can be influenced by contextual factors (Goldstein and Lackner, 1974; Morais and Landercy, 1977); whether the same is true with fused syllables remains to be investigated. As far as individual differences are concerned, automatic attentional effects are difficult to distinguish from the physiological or functional asymmetry itself; they are probably highly correlated. Strategic biases, on the other hand, are voluntary and at the disposition of the listener. For example, by deliberately paying attention to the left ear, even persons with a strong REA can produce a LEA with unfused stimuli. Such strategies are not under control in the standard two-response paradigm, so that the ear advantages obtained are not a pure measure of lateral asymmetry.

This is especially obvious in the single-response paradigm when applied to unfused stimuli. It does not suffice to instruct the listeners not to pay attention to either ear; especially inexperienced listeners may not follow these instructions, and there is no way of controlling whether they do. It may be difficult in principle to "neutralize" attention. Requiring two responses at least reduces the effect that attentional biases would have in the single-response paradigm. There remains the possibility of controlling the listener's strategies by instructions to pay attention to one or the other ear and to report only the stimuli in that ear. This procedure has been followed by several researchers, although usually not for the purpose of assessing ear advantages. It may be considered as a two-response paradigm in two passes; in this case, a single ear advantage index would be computed after combining the results of the two (properly counterbalanced) selective-attention conditions. The problem is that here, because of the relative efficiency of selective attention, performance level will be rather high, making the ear advantage index less reliable. Alternatively, the two selective-attention conditions may be considered as single-response paradigms, and two separate single-response indices may be computed whose difference is then taken as the measure of the ear advantage. However, here we encounter the same problem as with the d index: simple differences depend on the absolute size of the numbers involved, so that the resulting index reflects individual differences in the relative effectiveness of selective attention in addition to the ear advantage itself. Regardless of the form of data analysis, there is the theoretical possibility that there are lateral asymmetries in the effectiveness of voluntary selective attention that are independent of the ear advantage itself and again would confound the measure of the ear advantage.

We conclude that there is no perfect way of controlling attentional strategies with unfused stimuli, so that fused stimuli offer a significant methodological advantage in this respect. Future research will determine whether the relatively small ear advantages obtained with perfectly fused syllables are the relatively "pure" measure of physiological asymmetry that I

suspect them to be.

4.5. Blend Responses

In the discussion of the single-response paradigm (Section 3), it was assumed that only two kinds of responses are given to a fused dichotic pair; they match one of the two component stimuli and can be assigned to one or the other ear. Of the fifteen possible combinations of the six standard stop-consonant-vowel syllables, only seven meet this strict criterion, given that they are highly intelligible in isolation. These pairs are the place contrasts /ba/-/da/, /da/-/ga/, /pa/-/ta/ and /ta/-/ka/, and the voicing contrasts /ba/-/pa/, /da/-/ta/ and /ga/-/ka/. These are the stimulus pairs especially suited for the methodology outlined in Section 3.

However, it may be desirable for some purpose to include other stimulus combinations as well, and past experiments have almost always included all possible combinations of the stimuli. The two place contrasts, /ba/-/ga/ and /pa/-/ka/, may receive a third response, /da/ and /ta/, respectively. Cutting (1976) has called these intermediate percepts "psychoacoustic fusions." Their frequency may be negligible for many listeners, but some give a substantial proportion of these responses (Repp, 1976b). The remaining stimulus combinations are the six double-feature contrasts: /ba/-/ta/, /ba/-/ka/, /da/-/pa/, /da/-/ka/, /ga/-/pa/ and /ga/-/ta/. They typically yield two additional responses per pair, resulting from the combination of the feature values of the component stimuli; for example, /ba/-/ta/ is heard not only as /ba/ or /ta/, but also as /pa/ and /da/. These "blend" responses are usually quite frequent and may even exceed the proportions of correct responses, although there is much variation between stimulus pairs and subjects in this respect (Halwes, 1969; Repp, 1977a). Blend responses and psychoacoustic fusions usually do not convey direct information about ear asymmetries, so that the question arises what to do with them.

Hybrid responses also occur with unfused syllables (Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970). In the two-response paradigm, they are simply grouped together with other types of errors in the class of incorrect responses. As a result, double-feature contrasts typically have higher error rates than single-feature contrasts, an effect that has been termed the "feature-sharing advantage" (Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy et al., 1972; Pisoni, 1975). The availability of the correct-incorrect distinction makes it easy to dispose of blends in the two-response paradigm.

In the single-response paradigm, on the other hand, we have assumed that all responses are "correct," apart perhaps from a few random errors (that may be divided up between the two response categories). Blend responses are different from random errors in that they reflect what the listener actually heard; in a sense, they are correct responses. However, they cannot be unambiguously assigned to one or the other ear. There are two ways of dealing with them. One possibility, followed by Repp (1977a), is to analyze the data in terms of the individual phonetic features and to calculate two laterality indices, one for voicing and one for place. If only one feature is considered at a time, blend responses become informative with respect to lateral asymmetries. The two resulting indices may be averaged to obtain a single index.

The other possibility, which consists of discarding blend responses, is more problematic. Consider the following example, shown in Table 4:

TABLE 4: Fictitious response distribution for a double-feature contrast pair.

Stimuli		Responses			
LE	RE	/ba/	/ta/	/pa/	/da/
/ba/-	/ta/	0	.33	.33	.33
/ta/-	/ba/	.5	0	.25	.25

Here omission of blends (/pa/ and /da/ responses) would lead to the conclusion that there is a perfect REA for this stimulus pair. However, when the data are analyzed for each feature separately, it is found that there are only moderate REAs for voicing and place (e = +0.46 for both). If blend errors are discarded, this information is lost and the REA is inflated. It is not clear which should be considered the correct index: the average of the separate indices for the two features or the index based on "correct" responses only.

It may be possible to settle the problem by examining empirical isolaterality contours (ROC functions) for single- and double-feature contrast pairs. In the meantime, double-feature contrasts and pairs yielding psychoacoustic fusions are best omitted from dichotic single-response tests, as long as only the ear advantage is of concern. This leaves us with only seven of the original fifteen stimulus pairs -- perhaps too few to constitute a useful test. However, synthetic stimuli offer the possibility of varying the acoustic structure of the stimuli while leaving their phonetic content unchanged. By varying voice onset time or the formant transitions within phonetic categories, stimulus dominance relationships can be changed, so that an e' index can easily be calculated (Repp, 1976b, 1977a). In fact, it is possible to take a single stimulus pair (for example, /ba/-/pa/), to select several tokens with different acoustic characteristics (for example, four different voice onset times within each category), and thus to arrive at a test that contains a sufficient number of stimulus pairs (sixteen combinations), is maximally homogeneous, and leads to a clean estimate of ear dominance (see footnote 6a).[7] This illustrates one of the great methodological advantages of the single-response paradigm over the two-response paradigm; the latter always requires a larger number of response alternatives in order to reduce the effect of guessing.

[7] In principle, e' can be calculated without varying stimulus dominance. However, varying the stimuli and, with them, stimulus dominance is important in order to avoid extreme dominance asymmetries due to individual idiosyncrasies, to derive an ROC function, and simply to provide variety in the test.

4.6. Test Reliability

In the Introduction, I have stressed that dichotic tests must satisfy general test-theoretical standards. One of these is reliability. As in any other psychological test, the observed score (ear advantage) of a subject represents his or her "true" score plus random measurement error. The magnitude of the measurement error depends on the length of the test. It is not surprising that, in repeated administrations of a short dichotic test, the observed ear advantages for a given subject vary considerably and may even show reversals in direction (Speaks, Niccum, and Carney, 1976). Most dichotic studies in the past have used short tests whose reliability was likely to be low. The fact that a certain percentage of right-handed subjects show either no REA or an LEA (although physiological data suggest that virtually all are left-hemisphere-dominant for speech) is at least due in part to measurement error (cf. Blumstein, Goodglass, and Tartter, 1975).

Ryan and McNeil (1974) have reported a test-retest reliability coefficient of +0.80 for a 60-item test, and Blumstein et al. (1975) found a somewhat lower coefficient of +0.74 for an 80-item test. Both studies used natural-speech CV syllables in the two-response format. These reliabilities are quite satisfactory in view of the relative shortness of the tests and the weaknesses of the two-response paradigm (guessing, attentional fluctuations, etc.). Researchers in the field have tended to expect too much from a short dichotic test and have been reluctant to accept the conclusion that much longer tests will be necessary to obtain precise measurements. If we accept the Blumstein et al. results as typical and apply the standard Spearman-Brown formula (Lord and Novick, 1968, p. 112), we find that the test has to be three times as long (about 240 pairs) to achieve a reliability of +0.90, and six times as long (about 480 pairs) to reach $r = +0.95$. From the Ryan and McNeil data, we obtain more moderate estimates of 140 and 280 pairs, respectively. Considering the fact that the standard set of six CV syllables yields a basic test unit of 30 dichotic pairs, I would recommend that ten repetitions of this test unit (that is, 300 pairs) be administered in order to obtain stable ear advantage indices. Such a test requires about 20 minutes of listening time and therefore should be feasible under most circumstances, both in and outside the laboratory.

Underlying the development of the single-response methodology is the hope that this procedure will prove to be more reliable than the traditional two-response paradigm. Alternative methods, such as the AXB paradigm mentioned earlier, may also lead to increased reliability. I plan to conduct pertinent studies in the near future.[8]

4.7. Test Homogeneity and Validity

The problem of test reliability is a practical one that always can be solved by using a test of sufficient length. More important is the

---

[8] Extremely encouraging results have been obtained recently by Bruce Wexler (personal communication, 1976). Using a 60-item test of relatively unfused syllables in a single-response paradigm, Wexler obtained reliabilities well above +0.90 for both normal and psychotic subjects. (See also footnote 6a.)

theoretical question of what is actually being measured--the validity of the test. Ultimately, its validity as an instrument for assessing hemispheric dominance needs to be assessed by physiological criteria of functional lateralization. At present, however, these physiological measurements are still crude and hazardous; moreover, they are a less crucial criterion than they may seem at first thought. First of all, the only reliable physiological indicator in normal subjects, the Wada test, yields only categorical outcomes (left, right, or no dominance), not a graded scale of lateralization. Moreover, it really supplies a useful criterion only for the small group of left-handers, since it is now well-established that virtually all right-handers are left-hemisphere-dominant for speech. Secondly, the original idea that the dichotic ear advantage directly reflects hemispheric dominance for speech is probably an oversimplification. It is likely that there are multiple factors underlying the dichotic ear advantage, only one of which is the (quite possibly all-or-none) dominance of one hemisphere for speech. The primary task of the theoretical study of the ear advantage must therefore be to determine what is actually being measured. This is a difficult problem, but some preliminary steps are possible by asking the following familiar test-theoretical questions: Do all items in a test measure the same underlying variable(s)? And do different tests composed of items from the same general class (viz., those that tend to yield an average REA) measure the same underlying variable(s)?

These important (and closely related) questions about within-test (or item) homogeneity and between-test homogeneity (or validity) have been totally ignored in the past. Their answers are by no means obvious. Consider the question of item homogeneity. Repp (1976b), for example, found that two fused stimulus pairs of a three-item test yielded REAs but the third pair did not. More evidence on this problem is needed. The statistical techniques that may be applied are intercorrelation of laterality indices for individual items in a test and subsequent factor analysis, or perhaps an adaptation of the more recent methods of stochastic test theory (Rasch, 1960; Lord and Novick, 1968). These analyses should determine whether all items in the test measure a single factor, or whether different items measure different factors. The derivation of an ROC function in the single-response paradigm (Section 3.2) also constitutes a (less rigorous) test of item homogeneity. (However, even if it turned out that only a single factor is being measured, this would show only that the test is homogeneous and all items measure the same thing; the single factor may nevertheless represent a complex of underlying variables.)

A related problem is whether the ear advantages for different phonetic features reflect the same underlying factors. In a recent study of partially fused dichotic double-feature contrasts, I obtained e' coefficients separately for voicing and place; they correlated only +0.64, although each index was based on the same 768 trials (Repp, 1977a). I hypothesized that individual differences in perceptual organization may be reflected in the dichotic ear advantage (see also Section 4.1). Tests of this hypothesis are needed.

Finally, the homogeneity question needs to be asked about whole tests: Do tests composed of different types of speech stimuli (CVs, VCs, VCVs, or words; stops, fricatives, or nasals; etc.) measure the same factor? Do tests composed of natural-speech syllables measure the same factor as synthetic tests? Do fused and unfused syllables (or: the single-response and the two-

response paradigm) assess the same factor? Again, intercorrelations between different tests (perhaps supplemented by factor analysis and modern test theory) should provide an answer. So far, there are no data available. A positive result would reassure us that we are actually measuring a well-defined characteristic whose complexities will have to be unravelled primarily by physiologists. Negative results, on the other hand, disastrous as they would be for the diagnostic application of dichotic tests, would be of great theoretical interest. Perhaps there is more than one "ear advantage," that is, different tests may tap different dimensions of a very complex phenomenon.[9]

4.8. Absolute Magnitude of the Ear Advantage

The questions of reliability, homogeneity, and validity, which are correlational in nature, must be kept separate from the issue of the absolute magnitude of the ear advantage. For example, ear advantages may increase with practice (although the evidence appears to be negative--see Porter, Troendle and Berlin, 1976), but as long as they do so for all individuals, the reliability of the test will not be affected. Different items in a test may yield different magnitudes of ear advantages, but they nevertheless may measure the same underlying variable. Similarly, different classes of stimuli may yield different average magnitudes of REA and nevertheless measure the same thing. As long as all individuals tested are in basically the same rank order on each test (or on each item), the homogeneity criterion is satisfied, and it is immaterial which tests or items are selected for testing persons, as long as all persons to be compared are tested with the same tests or items. The variations in the absolute magnitude of the ear advantage represent variations in item or test "difficulty," in terms of test theory. It is a separate but nevertheless important question what causes these variations in difficulty, if they exist. On the other hand, if two items or tests yield the same average REA, this implies absolutely nothing about their intercorrelation.

One striking difference in the magnitude of ear advantages has been discovered in recent research using the single-response paradigm (see Repp, 1976b, 1976c, 1977a, and footnote 6a): partially fused syllables (voicing and double-feature contrasts) yield much larger ear advantages than perfectly fused syllables (place contrasts). This result is methodologically interesting, because larger ear advantages are also likely to be more reliable. The reason for this difference is not clear at present, except that perfect fusion seems to play a role. The role of selective attention with partially fused stimuli needs to be reassessed, although earlier studies suggest that it is small (Halwes,1969; Repp, 1976a). Future research will concentrate on determining the factors that are responsible for this difference between fused and partially fused syllables.

---

[9] I am referring here to tests at the same level of complexity, varying only in the auditory and phonetic properties of the stimuli. There is good reason to believe that dichotic tasks of different complexity tap different aspects of lateralization (Porter and Berlin, 1975).

REFERENCES

Berlin, C. I., Lowe-Bell, S. S., Cullen, J. K., Jr., Thompson, C. L., and Loovis, C. F. (1973) Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. J. Acoust. Soc. Am. 53, 699-709.

Blumstein, S., Goodglass, H., and Tartter, V. (1975) The reliability of ear advantage in dichotic listening. Brain Lang., 2, 226-236.

Broadbent, D. E. (1955) A note on binaural fusion. Quart. J. Exp. Psychol., 7, 46-47.

Broadbent, D. E. and Ladefoged, P. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Am. 29, 708-710.

Cullen, J. K., Jr., Thompson, C. L., Hughes, L. F., Berlin, C. I., and Samson, D. S. (1974) The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain Lang., 1, 307-322.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychol. Rev., 83, 114-140.

Franklin, B. (1969) The effect on consonant discrimination of combining a low-frequency passband in one ear and a high-frequency passband in the other ear. J. Audit. Res., 9, 365-378.

Goldstein, L. and Lackner, J. R. (1974) Sideways look at dichotic listening. J. Acoust. Soc. Am. 55 (Supplement), S10(A).

Green, D. M. and Swets, J. A. (1966) Signal Detection and Psychophysics. (New York: Wiley).

Haggard, M. P. (1975) Asymmetrical analysis of stimuli with dichotically split formant information. Speech Perception, Series 2, No. 4, 11-19. (Dept. of Psychology, The Queen's University of Belfast.)

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Supplement to Haskins Laboratories Status Report on Speech Research.

Harshman, R. and Krashen, S. (1972) An "unbiased" procedure for comparing degree of lateralization of dichotically presented stimuli. UCLA Working Papers in Phonetics, 23, 3-12.

Hublet, C., Morais, J., and Bertelson, P. (1976) Spatial constraints on focused attention: Beyond the right-side advantage. Perception, 5, 3-8.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol., 15, 166-171.

Kinsbourne, M. (1973) The control of attention by interaction between the cerebral hemispheres. In Attention and Performance IV, ed. by S. Kornblum. (New York: Academic), pp. 239-256.

Kinsbourne, M. (1975) The mechanism of hemispheric control of the lateral gradient of attention. In Attention and Performance V, ed. by P. M. A. Rabbit and S. Dornic. (London: Academic), pp. 81-97.

Kuhn, G. M. (1973) The phi coefficient as an index of ear differences in dichotic listening. Cortex, 9, 447-457.

Leakey, D. M., Sayers, B. McA., and Cherry, C. (1958) Binaural fusion of low-and high-frequency sounds. J. Acoust. Soc. Am., 30, 222.

Levy, J. (in press) The correlation of the (phi) function difference score with performance. Cortex.

Lord, F. M. and Novick, M. R. (1968) Statistical Theories of Mental Test Scores. (Reading, Ma: Addison-Wesley).

Marshall, J. C., Caplan, D., and Holmes, J. M. (1975) The measure of

laterality. Neuropsychologia, 13, 315-321.
Mills, A. W. (1972) Auditory localization. In Foundations of Modern Auditory Theory, vol. II, ed. by J. V. Tobias. (New York: Academic), pp. 301-348.
Morais, J. (1975) The effects of ventriloquism on the right side advantage for verbal material. Cognition, 3, 127-139.
Morais, J. and Bertelson, P. (1973) Laterality effects in diotic listening. Perception, 2, 107-111.
Morais, J. and Bertelson, P. (1975) Spatial position versus ear of entry as determinant of the auditory laterality effects: A stereophonic test. J. Exp. Psychol.: Human Percep. Perform., 1, 253-262.
Morais, J. and Landercy, M. (1977) Listening to speech while retaining music: What happens to the right-ear advantage? Brain Lang., 4, 295-308.
Nearey, T. M. and Levitt, A. G. (1974) Evidence for spectral fusion in dichotic release from upward spread of masking. Haskins Laboratories Status Report on Speech Research, SR-39/40, 81-89.
Niccum, N., Speaks, C., and Carney, E. (1976) Reversals in ear advantage with dichotic listening: Effects of alignment. J. Acoust. Soc. Am., 59S, S6(A).
Nye, P., Nearey, T., and Rand, T. (1974) Dichotic release from masking: Further results from studies with synthetic speech stimuli. Haskins Laboratories Status Report on Speech Research, SR-37/38, 123-137.
Odenthal, D. W. (1963) Perception and neural representation of simultaneous dichotic pure tone stimuli. Acta Physiologica et Pharmacologica Neerlandica, 12, 453-496.
Perrott, D. R. (1970) Signal and interaural level difference effects on binaural critical band. J. Audit. Res., 10, 1-4.
Perrott, D. R. and Barry, S. H. (1969) Binaural fusion. J. Audit. Res., 3, 263-269.
Perrott, D. R., Briggs, R., and Perrott, S. (1970) Binaural fusion: Its limits as defined by signal duration and signal onset. J. Acoust. Soc. Am., 47, 565-568.
Pisoni, D. B. (1975) Dichotic listening and processing phonetic features. In Cognitive Theory, vol. I, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
Porter, R. J., Jr., and Berlin, C. I. (1975) On interpreting developmental changes in the dichotic right-ear advantage. Brain Lang., 1975, 2, 186-200.
Porter, R. J., Jr., Troendle, F., and Berlin, C. I. (1976) Effects of practice on the perception of dichotically presented stop-consonant-vowel syllables. J. Acoust. Soc. Am., 59, 679-682.
Preston, M. S., Yeni-Komshian, G., and Benson, P. (1968) A dichotic ABX procedure for use in laterality studies. Haskins Laboratories Status Report on Speech Research, SR-13/14, 179-180.
Rand, T. C. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Am., 55, 678-680.
Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. (Copenhagen: Nielson and Lydiche).
Repp, B. H. (1976a) Effects of fundamental frequency contrast on discrimination and identification of dichotic CV syllables at various temporal delays. Mem. Cog., 4, 75-90.
Repp, B. H. (1976b) Identification of dichotic fusions. J. Acoust.

Soc. Am., 60, 456-469.
Repp, B. H. (1976c) Discrimination of dichotic fusions. Haskins Laboratories Status Report on Speech Research, SR-45/46, 123-139.
Repp, B. H. (1977a) Dichotic competition of speech sounds: The role of acoustic stimulus structure. J. Exp. Psychol.: Human Percep. Perform., 3, 37-50.
Repp, B. H. (1977b) A simple model of response selection in the dichotic two-response paradigm. Haskins Laboratories Status Report on Speech Research, SR-49.
Richardson, J. T. E. (1972) Nonparametric indexes of sensitivity and response bias. Psychol. Bull., 78, 429-432.
Ryan, W. I. and McNeil, M. (1974) Listener reliability for a dichotic task. J. Acoust. Soc. Am, 56, 1922-1923.
Shankweiler, D. and Studdert-Kennedy, M. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol., 19, 59-63.
Shankweiler, D. and Studdert-Kennedy, M. (1975) A continuum of lateralization for speech perception? Brain Lang., 2, 212-225.
Speaks, C., Niccum, N., and Carney, E. (1976) Noninvariance of the ear advantage in dichotic listening. J. Acoust. Soc. Am., 60, S119(A).
Speaks, C., Niccum, N., Carney, E., and Marble, K. (1975) Dichotic listening: Reversals in ear advantage. J. Acoust. Soc. Am., 58, S76(A).
Studdert-Kennedy M. and Shankweiler, D. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Am, 48, 579-594.
Studdert-Kennedy, M., Shankweiler, D., and Pisoni, D. B. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol., 3, 455-466.
Tobias, J. V. (1972) Curious binaural phenomena. In Foundations of Modern Auditory Theory (vol. II), ed. by J. V. Tobias. (New York: Academic), pp. 463-486.
Van den Brink, G., Sintnicolaas, K., and Van Stam, W. S. (1976) Dichotic pitch fusion. J. Acoust. Soc. Am, 59, 1471-1476.
Zangwill, O. L. (1960) Cerebral Dominance in Relation to Psychological Function. (Edinburgh: Oliver and Boyd).

A Simple Model of Response Selection in the Dichotic Two-Response Paradigm

Bruno H. Repp

ABSTRACT

A simple random-guessing model of the dichotic two-response paradigm is described. The model provides a way of calculating an index of the ear advantage that takes guessing into account. It also generates predictions of the proportions of single- and double-correct responses at different performance levels. A comparison with real data shows that the proportions of double-correct responses are generally overpredicted. By introducing an additional parameter reflecting limited channel capacity, the model can be made to fit closer to empirical data, but the value of the parameter is not the same for different sets of data. While this model is oversimplified in many ways, it nevertheless provides a rudimentary formal framework for the interpretation of dichotic data.

INTRODUCTION

Despite a large amount of research and theoretical speculations on dichotic listening, little thought has been given to formulating and testing mathematical models of the response processes involved. The present paper briefly examines the simplest conceivable formal model and derives some predictions from it. The model is almost certainly an oversimplification. However, the purpose of the exercise is to point out some basic relations between several dependent variables in dichotic listening experiments. These relations are likely to hold up approximately, even if the model that predicts them is not precisely true, and they need to be taken into account in the interpretation of dichotic data.

The present paper serves as an appendix to my methodological paper, "Measuring Laterality Effects in Dichotic Listening" (Repp, 1977), to which frequent reference will be made.

AN INDEPENDENT-CHANNELS MODEL WITH RANDOM GUESSING

In Section 1 of the preceding paper, I discussed the dichotic two-response paradigm. This is the standard procedure that requires the listener to identify both stimuli on each trial. The two responses must be different from each other and are scored without regard to order. The proportions of correct responses for the right and left ear are P_R and P_L, respectively, and

---

Acknowledgment: Preparation of this paper was supported by NICHD Grant HD01994 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research (SR-49) (1977)]

the overall performance level is $P_o = (P_R + P_L)/2$. I described what is probably the best index of ear asymmetry (I called it e; it is basically identical with the f index of Marshall, Caplan, and Holmes, 1975), and I derived a correction for guessing. This correction is rather crude, based on linear interpolation between three extreme cases. I pointed out that a formal model of guessing would provide a more elegant solution.

The simplest model of response selection in the dichotic two-response paradigm makes the following two assumptions: (1) the stimuli in the two ears are perceived independently of each other; (2) a stimulus is either perceived correctly or a random guess is made. Although the real situation is almost certainly more complex, the predictions of such a simple model are worth considering. If P_R^* and P_L^* are the "true" probabilities of correctly perceiving the stimuli in the respective ears and N is the number of stimuli in the experiment, then

$$(1) \qquad P_R = P_R^* + (1 - P_R^*)P_L^*[1/(N - 1)] + (1 - P_R^*)(1 - P_L^*)(2/N) \text{ , and}$$

$$(2) \qquad P_L = P_L^* + (1 - P_L^*)P_R^*[1/(N - 1)] + (1 - P_L^*)(1 - P_R^*)(2/N) \text{ .}$$

The three additive terms in these equations are: (1) the probability of correctly perceiving the stimulus in the ear concerned; (2) the probability of making a correct guess when the stimulus in the other ear is correctly identified; and (3) the probability of making a correct guess when no stimulus is perceived correctly.

By taking the difference between these two equations, we find that

$$(3) \qquad \begin{aligned} d = P_R - P_L &= P_R^* - P_L^* - [1/(N - 1)](P_R^* - P_L^*) \\ &= [(N - 2)/(N - 1)](P_R^* - P_L^*) \\ &= [(N - 2)/(N - 1)]d^* \text{ .} \end{aligned}$$

Thus, the observed ear difference d is in a simple proportional relationship to the underlying ear difference d^*. The proportionality factor is identical with the largest possible expected d, d_{max}, for a given N (Repp, 1977: Eq. 15). This becomes obvious by noting that if $d = d_{max}$ then necessarily $d^* = d^*_{max} = d/d_{max} = 1$.

The numerical solution of Equations 1 and 2 for P_R^* and P_L^* is not straightforward, so that it will not be derived here. (The solution is found most easily by a recursive procedure.) After estimates of P_R^* and P_L^* are obtained, an appropriate index of the ear advantage is

$$(4) \qquad \begin{aligned} e^* &= (P_R^* - P_L^*)/(P_R^* + P_L^*) && \text{if } 0.0 \leq P_o^* \leq 0.5 \\ &= (P_R^* - P_L^*)/(2 - P_R^* - P_L^*) && \text{if } 0.5 \leq P_o^* \leq 1.0 \text{ .} \end{aligned}$$

Of course, Equation 4 is identical with the formula for the e index before the correction for guessing (Repp, 1977: Eq. 12), except that observed scores P_R and P_L are replaced by underlying probabilities P_R^* and P_L^* that are already corrected for guessing. One might expect e_g (the e index after the correction for guessing proposed in Repp, 1977: Eq. 20) to be identical with e^*, but this is not quite true, as illustrated in Figure 1.

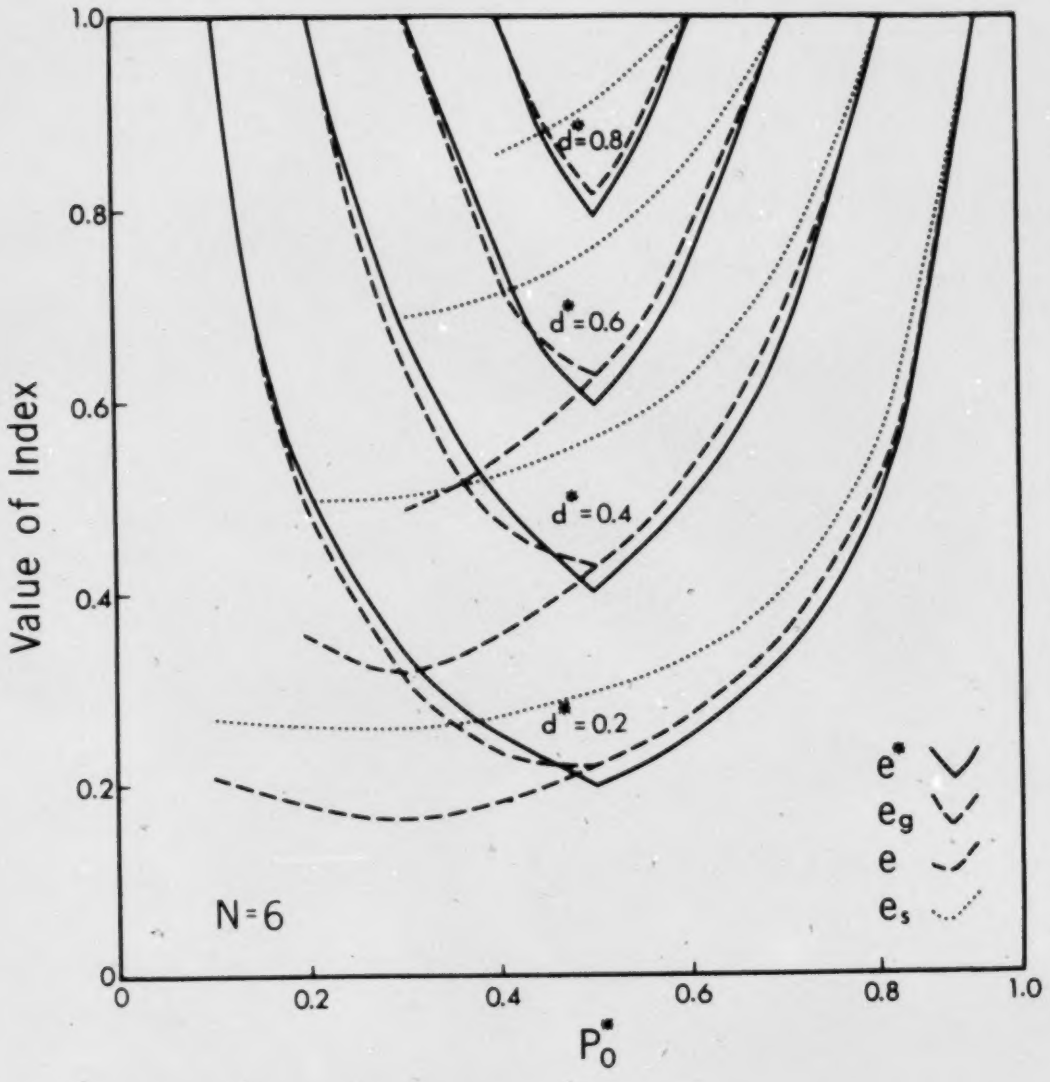


Figure 1: Four different indices of the ear advantage as a function of P_o^* and d^*.

Figure 1 shows four laterality indices as a function of two variables: P_o^* and d^*--the average and the difference, respectively, of the two underlying probabilities P_R^* and P_L^*. These are <u>not</u> isolaterality contours,

which are horizontal straight lines in Figure 1. Rather, each curve describes the value of the relevant index for a constant d^*, so they are "isodifference" contours.

The index e was proposed by Halwes (1969) and Marshall et al. (1975) without a correction for guessing (Repp, 1977:Eq. 12), whereas e_g incorporates the correction for guessing proposed in Repp (1977:Eq. 20). This correction has the effect of bending the left parts of the e functions ($P_o^* < 0.5$) upward, so that the functions become U-shaped and nearly symmetric. However, they are not perfectly symmetric, as the e^* functions are. The reason for this will become clear in the next section. Here we note only that e_g and e^* are nearly identical, which shows that the rough correction for guessing proposed earlier is compatible with the simple guessing model discussed here. Therefore, this correction should suffice for all practical purposes, and it generally will not be necessary to actually compute e^*.

The fourth index, e_S, was not discussed in Repp (1977) but deserves a brief comment here. Studdert-Kennedy and Shankweiler (1970) proposed to consider only single-correct trials for computing an index of the ear advantage, since double-correct responses provide no information about ear asymmetry. If P_{RS} and P_{LS} are the proportions of single-corrects for the two ears, and $P_S = P_{RS} + P_{LS}$, then P_{RS}/P_S constitutes an index of the ear advantage; or, alternatively,

$$(5) \qquad e_S = (P_{RS} - P_{LS})/P_S \ ,$$

is an equivalent index that ranges from -1 to +1. This index is plotted as a function of P_o^* in Figure 1, together with the other indices discussed earlier. The simple guessing model provides a useful theoretical comparison of different indices.

First of all, e_S obviously needs a correction for guessing that still needs to be derived. Secondly, e_S is clearly different from e, leading to larger values throughout. This is not necessarily an argument against e_S; as long as two indices are perfectly correlated (as they seem to be), one is as good as the other for ordinal measurement. The index e_S is based on increasingly fewer observations as P_o increases, so that its variability increases and its reliability decreases; however, the same is true for e. Thus, it seems that, with an appropriate correction for guessing, e_S would be an acceptable alternative to e_g or e^*. On the other hand, however, there is no reason why e_S should be used instead of e, for which the correction for guessing has been worked out and which is just as easy to compute. Certainly e^* and e_S indices are not directly comparable because they represent different scales of the ear advantage. Therefore, to maintain uniformity and comparability from study to study, e_S is not recommended for general use.

SINGLE- AND DOUBLE-CORRECT RESPONSES

Both e_g and e^* presuppose the validity of the model outlined in the first section. It is important to determine to which degree this model actually fits real data. One way of testing it consists in examining its

predictions of the proportions of double-correct and single-correct responses at different performance levels.

The proportion of double-correct responses, P_D, predicted by the simple guessing model is

$$(6) \quad P_D = P_R^* P_L^* + (1 - P_R^*)P_L^*[1/(N - 1)] + P_R^*(1 - P_L^*)[1/(N - 1)] + \\ + (1 - P_R^*)(1 - P_L^*)[2/N(N - 1)] .$$

The proportion of single-correct responses, P_S, is

$$(7) \quad P_S = P_R^*(1 - P_L^*)(N - 2)/(N - 1) + P_L^*(1 - P_R^*)(N - 2)/(N - 1) + \\ + (1 - P_R^*)(1 - P_L^*)(2/N) .$$

Alternatively, P_S may be obtained by subtraction:

$$(8) \quad P_S = P_{RS} + P_{LS} = (P_R - P_D) + (P_L - P_D) = P_R + P_L - 2P_D .$$

Of course, the overall performance level is

$$(9) \quad P_o = (P_R + P_L)/2 = P_S/2 + P_D .$$

Figure 2 shows P_D, P_S, and P_o as a function of P_o^* for the special case of N = 6. For each dependent variable, two functions are shown. One is curvilinear and represents the case of no ear advantage, $d^* = 0$. The other consists of two linear segments and represents the case of the maximal underlying ear difference at a given P_o^*, $d^* = d^*_{max}$ (given P_o^*) (cf. Repp, 1977: Eq. 3). The functions for constant values of d^* between 0 and d^*_{max} (given P_o^*) fall between the two extremes shown in Figure 2 and are parallel to the curvilinear function. The differences in proportions brought about by an increase in d^* from $d^* = 0$ is shown in detail in Figure 3. Here it can be seen more clearly that not only P_D and P_S, but also P_o depend on d^* (and, hence, on d), as well as on P_o^*. Thus, the observed performance level P_o is not completely independent of the observed ear difference d; according to the model, there is a slight negative relationship. This is the reason why e_g and e^* do not coincide in Figure 1. Only e^*, which is directly based on the model, takes the interdependence of P_o and d into account. However, while this effect is interesting from a theoretical viewpoint, it is negligible for practical purposes.

The P_D and P_S functions are of special interest. From Figure 2, it can be seen that, as performance level increases from chance, both P_D and P_S increase at first, but soon P_S begins to decrease rapidly while P_D continues to increase steadily. Figures 2 and 3 permit comparisons with real data. If the observed proportions P_o, d, P_S, and P_D are known, predicted proportions P_S and P_D are found as follows: first, d^* is determined from Equation 3; then P_o is located in Figure 2 on an interpolated function appropriate for d^* (the effect of d^* is so small that it may be ignored); then, the values of P_S and P_D for this P_o are determined on the ordinate in Figure 2; finally, P_S and P_D are corrected for the effect of d^* by Figure 3.

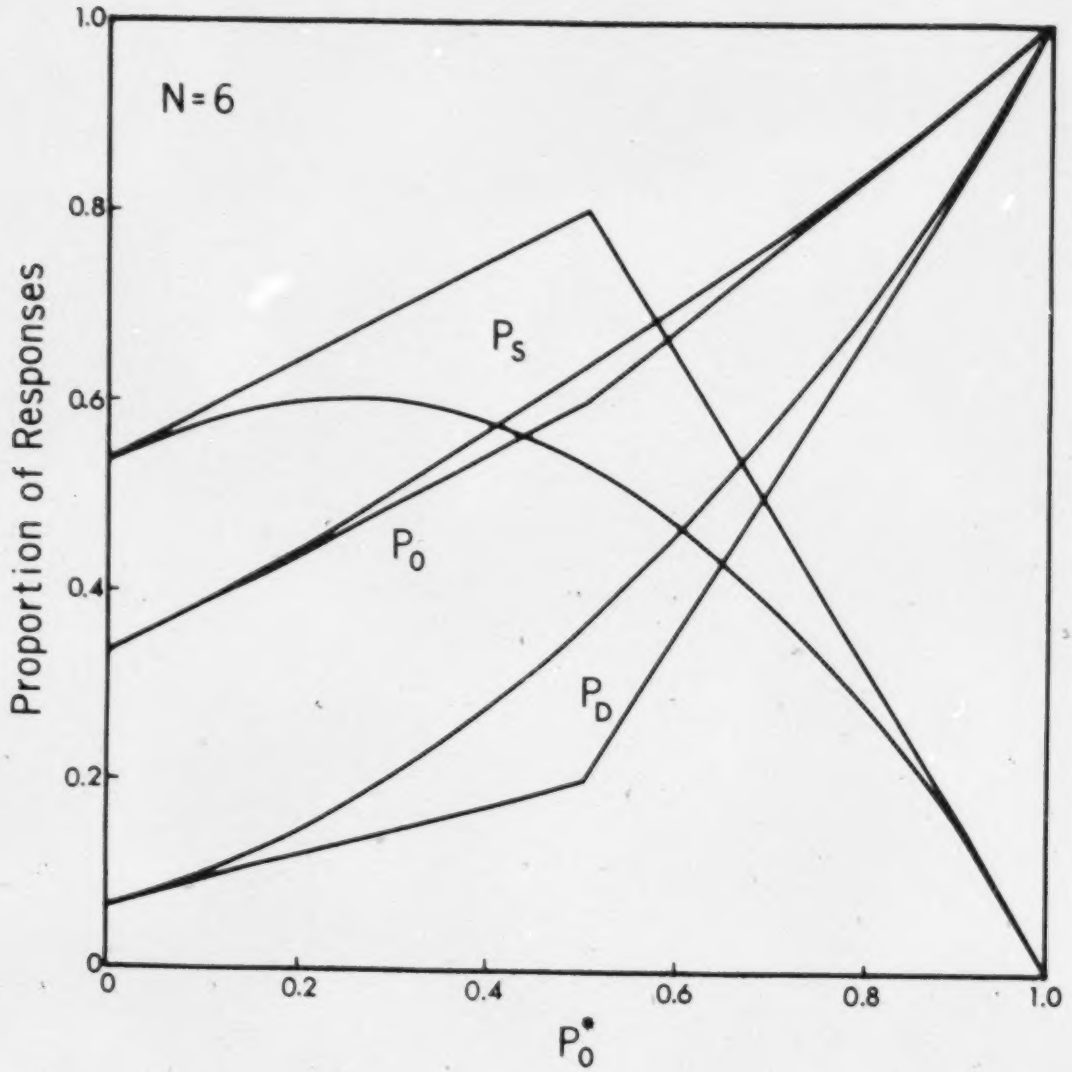


Figure 2: P_O, P_S, and P_D as a function of P_O^* and d^*.

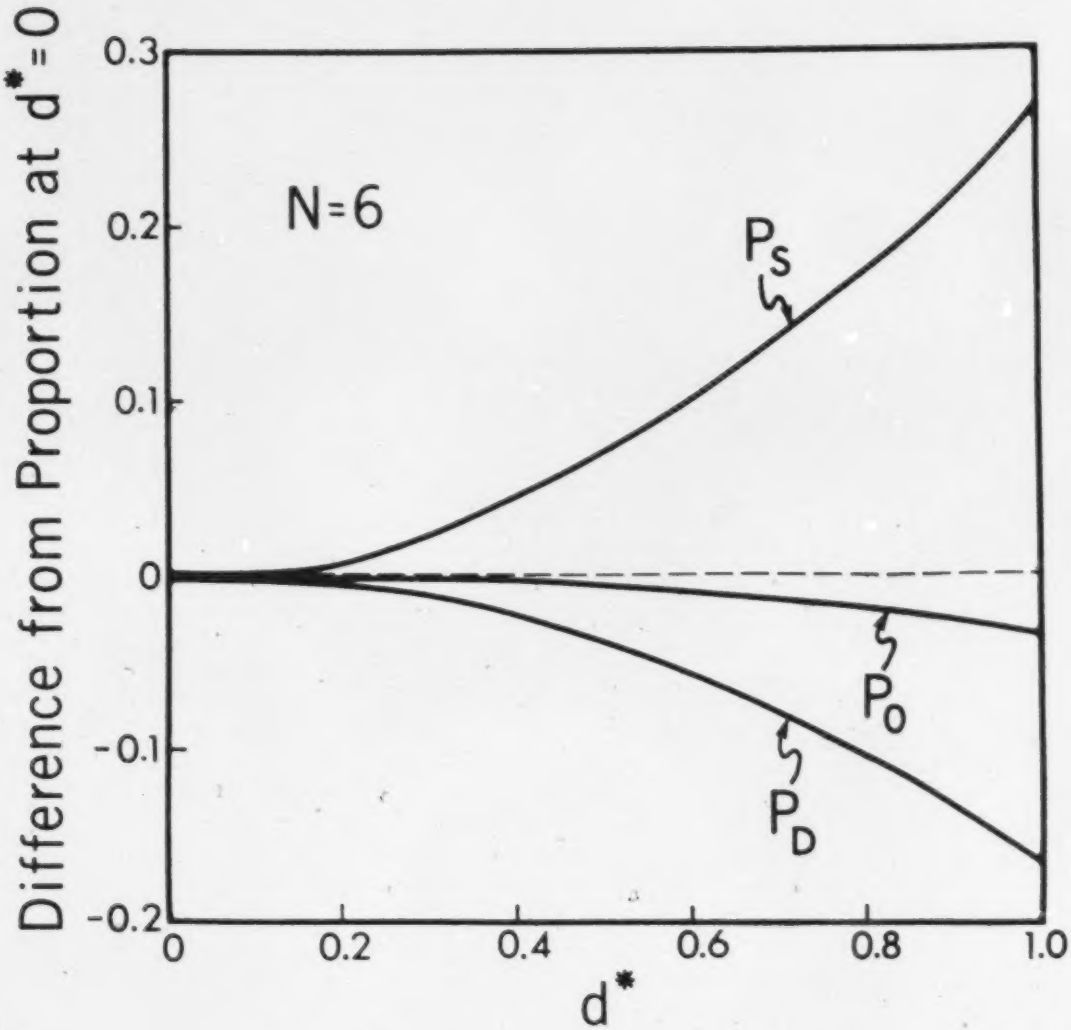


Figure 3: P_O, P_S, and P_D as a function of d^*.

TABLE 1: Comparison of some reported proportions of double- and single-correct responses with the predictions of the simple guessing model.

	P_o	d^*	P_S	P_D	$\hat{P}_S$	$\hat{P}_D$
A: Initial consonants	0.68	0.15	0.50	0.43	0.49	0.43
A: Medial vowels	0.82	0.03	0.28	0.69	0.32	0.66
A: Final consonants	0.74	0.08	0.44	0.52	0.42	0.53
B: 30 dB	0.52	0.15	0.66	0.19	0.58	0.22
B: 40 dB	0.62	0.09	0.59	0.32	0.54	0.35
B: 50 dB	0.71	0.15	0.52	0.45	0.46	0.48
C: Before practice	0.61	0.21	0.65	0.28	0.56	0.32
C: After practice	0.67	0.25	0.58	0.38	0.52	0.40
D: Clinical	0.37	0.13	0.62	0.06	0.56	0.09
D: Normal	0.50	0.14	0.67	0.18	0.60	0.20
E: 5 years	0.52	0.08	0.68	0.18	0.60	0.22
E: 7 years	0.56	0.10	0.68	0.22	0.58	0.27
E: 9 years	0.58	0.11	0.68	0.24	0.57	0.30
E: 11 years	0.61	0.13	0.67	0.27	0.55	0.33
E: 13 years	0.61	0.14	0.65	0.28	0.55	0.33

A = Studdert-Kennedy and Shankweiler (1970)
B = Cullen et al. (1974)
C = Porter et al. (1976)
D = Tobey et al. (1976)
E = Berlin et al. (1973)

Note: All studies used natural speech and the six stop consonants preceding /a/ (study A used CVC utterances).

Several studies in the literature report all the necessary parameters for several experimental conditions with different performance levels. These data and the predictions from the model are shown in Table 1. It can be seen that the model overpredicts P_D and underpredicts P_S in all cases but one. In other words, the observed proportions of double-correct responses are consistently smaller than predicted by the model. This indicates a negative dependency between P_R^* and P_L^*, such that the probability of perceiving the stimulus in one ear correctly is reduced if the stimulus in the other ear has already been perceived correctly. This effect is plausible in view of factors like fusion, selective attention, and memory, all of which tend to reduce perceptual accuracy for one channel to the degree that they increase accuracy for the other.

NONINDEPENDENCE OF CHANNELS

The model represented by Equations 1 and 2 assumed that errors in dichotic performance arise only from a very general form of processing limitation that reduces accuracy for both ears relative to monaural perfor-

mance, but permits independent perception of the degraded stimuli in each ear (cf. the "perceptual noise" hypothesis of Repp, 1975a, 1975b). However, the relatively poor fit of the model indicates that this assumption is not sufficient. Apparently, there is, in addition, a more specific processing limitation that makes it difficult to identify a second stimulus after one stimulus has been correctly perceived. (This is one way of conceptualizing the problem.) This limitation can be easily modelled by introducing one additional parameter into the model. Let us assume that the conditional probability of perceiving the stimulus in one ear correctly, given that the stimulus in the other ear has already been correctly identified, is reduced by a multiplicative factor c with respect to the same probability, given that the stimulus in the other ear has not been correctly identified. Thus,

(10) P_R^* L correct = cP_R^* L not correct, and

(11) P_L^* R correct = cP_L^* R not correct .

The constant c varies between 0 and 1; c = 0 indicates that, if the stimulus in one ear is correctly identified, the other stimulus can never be correctly identified except by a random guess; c = 1 indicates complete independence of the two channels. The full model, stated in terms of P_D and P_S, that will now be called P_D' and P_S', respectively, is:

$$(12) \quad P_D' = cP_R^*P_L^* + [P_R^*(1 - cP_L^*) + P_L^*(1 - cP_R^*) + (1 - P_R^*)P_L^* + (1 - P_L^*)P_R^*]/2(N - 1) + (1 - P_R^*)(1 - P_L^*)[2/N(N - 1)] ,$$

and

$$(13) \quad P_S' = [P_R^*(1 - cP_L^*) + P_L^*(1 - cP_R^*) + (1 - P_R^*)P_L^* + (1 - P_L^*)P_R^*][(N - 2)/2(N - 1)] + (1 - P_R^*)(1 - P_L^*)(2/N) .$$

In this version of the model, it makes a difference which channel is processed first; this results in the additional terms in the equations and in the additional "2" in the numerator. The simplifying assumption needs to be made that each channel is equally likely to be processed first, so that ear differences rest solely on differences between P_R^* and P_L^*. (Relaxing this assumption would lead to a more complex model that cannot be considered here.)

The effect of a decrease in c from 1 towards 0 is best illustrated by the differences between P_D and P_D' and between P_S and P_S'. Subtracting Equation 12 from Equation 6, one obtains after some rearrangement of terms,

$$(14) \quad P_D - P_D' = P_R^*P_L^*(1 - c)(N - 2)/(N - 1) ,$$

and subtracting Equation 13 from Equation 7,

$$(15) \quad P_S - P_S' = -P_R^*P_L^*(1 - c)(N - 2)/(N - 1) .$$

Thus, P_S and P_D change in precisely complementary fashion, resulting in a decrease in P_o as c decreases (cf. Equation 9). This is illustrated in Figure 4 that shows P_o', P_D', and P_S' as a function of P_o^* for three values of c: 0, .5, and 1. It is assumed that N = 6 and $d^* = P_R^* - P_L^* = 0$. Since

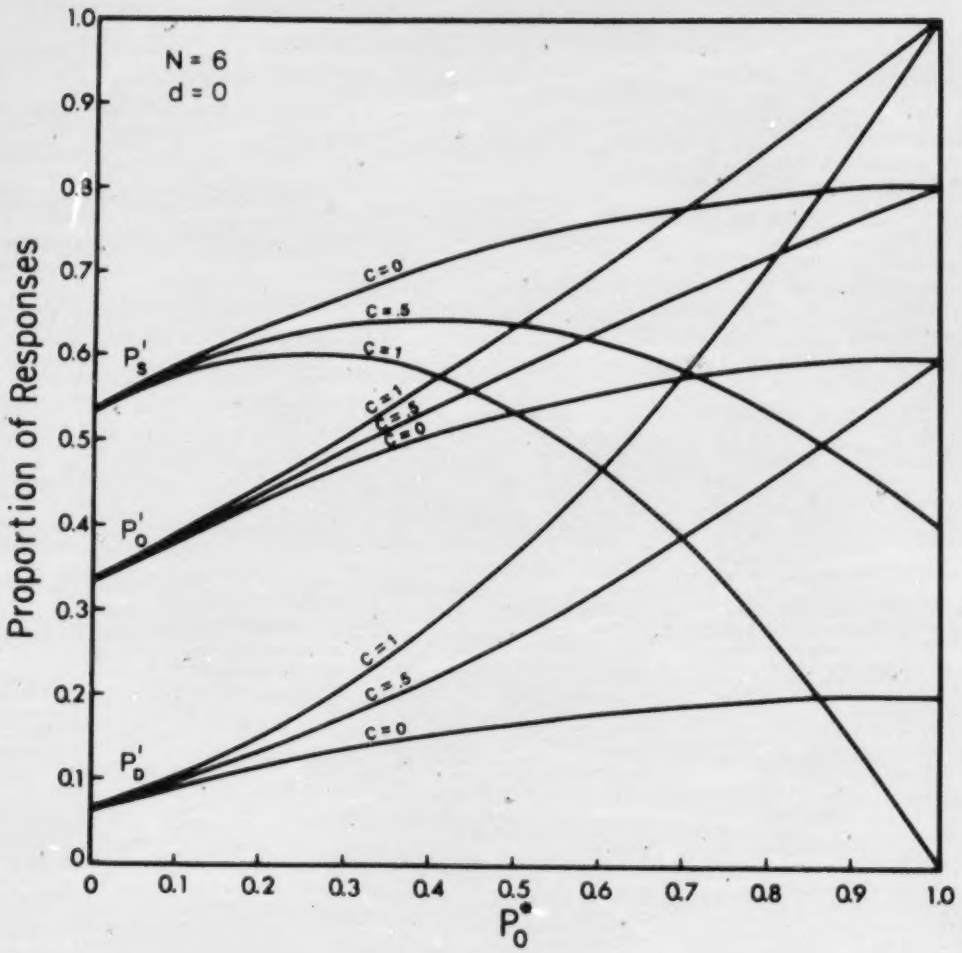


Figure 4: P_O, P_S, and P_D as a function of P_O^* and c.

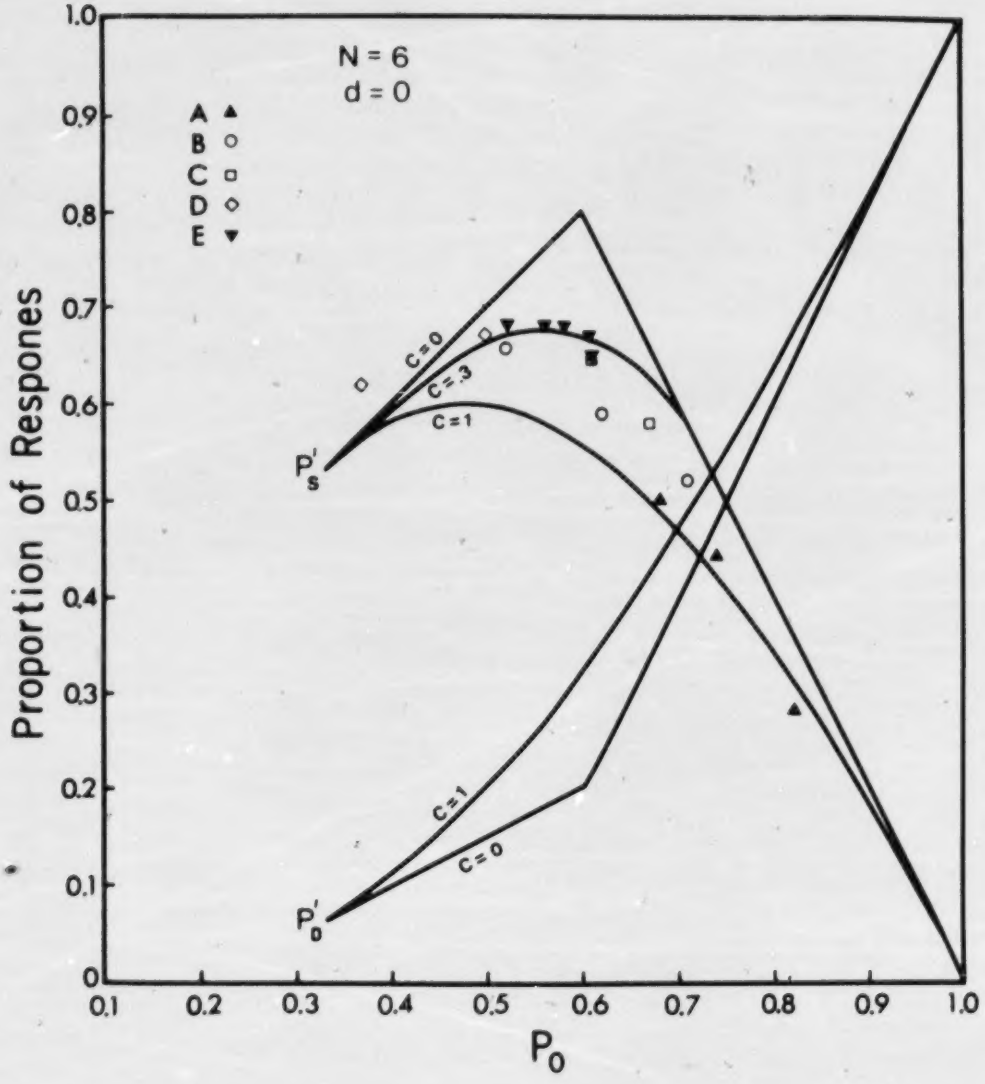


Figure 5: P_S and P_D as a function of P_O and c, compared with the data from Table 1.

and Table 1), the effect of ear differences on these functions may be neglected with little loss in accuracy (cf. Figure 3). The functions for c = 1 in Figure 4 are identical with the curvilinear functions in Figure 2. It can be seen that complete negative dependency between the two channels (c = 0) reduces the maximal expected performance level to 0.6 and the maximal expected proportion of double-correct responses to 0.2.

The model is compared to the data of Table 1 in Figure 5. The P_D' and P_S' functions of Figure 4 for c = 0 and c = 1 have been replotted here with P_o, the expected or observed performance level on the abscissa, in order to facilitate comparisons with real data. The long curvilinear functions are for c = 1, the short linear functions for c = 0. Functions with c between 0 and 1 lie between these two extremes, starting at the same point at the left and extending up to a point on the long linear segments that represent the maximal expected scores for different values of c. Only the observed P_S scores from Table 1 are plotted. (The differences between observed and predicted P_S scores are exactly twice as large as those between observed and predicted P_D scores, and therefore make discrepancies easier to see.)

Either of two conclusions can be drawn from Figure 5. If all data points are to be fit by a single function (and it seems that they could be), then the model is incorrect, for it cannot generate this function. On the other hand, it is possible that different experiments, stimuli, or groups of subjects require different functions. The three data points of study A (Studdert-Kennedy and Shankweiler, 1970) are fit by a function with c = 1, indicating virtual independence of channels. Eight of the other twelve data points seem to be fit by a function with approximately c = .3 (that has been drawn in Figure 5), indicating substantial negative dependencies between channels. The other data points require intermediate values of c, except for one point that falls completely outside the range of the model. Variations in c as a function of stimuli or subjects are not implausible. The stimuli of study A, for example, were different in several ways from those of studies B - E, which all come from the same laboratory. In this case, c may serve as an indicator of the degree of channel interaction (for example, fusion) in an experiment.

While further research will be required to evaluate the usefulness of the present model for making global predictions, it is clear that the model is not sufficient at a detailed level of analysis. For example, it could not explain stimulus dominance or the feature-sharing effect (see Repp, 1977). However, its gross predictions are likely to be not too far from the truth. The model has implications for researchers who have focused on P_D as a possible indicator of auditory processing capacity that is semi-independent of overall performance level (Berlin, Hughes, Lowe-Bell, and Berlin, 1973; Dermody and Noffsinger, 1976; Tobey, Cullen, and Rampp, 1976). The results of two such studies are included in Table 1 and in Figure 5. Tobey et al. (1976) noted that their two groups of subjects (children with and without auditory processing disorders) did not differ in P_S, but only in P_D. Similarly, Berlin et al. (1973) found that P_D increased with age, while P_S decreased, but to a much lesser extent. As can be seen in Table 1 and the figures, both findings are predicted by the present model. The subjects in both studies performed at relatively low levels, where P_S is nearly constant with changes in performance level. Therefore, the findings should be

ascribed to changes in overall performance level, not to any specific factor reflected by P_D alone.

REFERENCES

Berlin, C. I., Hughes, L. F., Lowe-Bell, S. S., and Berlin, H. L. (1973) Dichotic right ear advantage in children 5 to 13. Cortex 9, 394-402.

Cullen, J. K., Jr., Thompson, C. L., Hughes, L. F., Berlin, C. I., and Samson, D. S. (1974) The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain Lang. 1, 307-322.

Dermody, P. and Noffsinger, P. D. (1976) Auditory processing factors in dichotic CV tasks. J. Acoust. Soc. Am. 59 (Supplement), S6 (A).

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. dissertation, University of Minnesota. [Also supplement to Haskins Laboratories Status Report on Speech Research, September, 1969.]

Marshall, J. C., Caplan, D., and Holmes, J. M. (1975) The measure of laterality. Neuropsychologia 13, 315-321.

Porter, R. J., Jr., Troendle, R., and Berlin, C. I. (1976) Effects of practice on the perception of dichotically presented stop-consonant-vowel syllables. J. Acoust. Soc. Am. 59, 679-682.

Repp, B. H. (1975a) Dichotic forward and backward "masking" between CV syllables. J. Acoust. Soc. Am. 57, 483-496.

Repp, B. H. (1975b) Distinctive features, dichotic competition, and the encoding of stop consonants. Percep. Psychophys. 17, 231-240.

Repp, B. H. (1977) Measuring laterality effects in dichotic listening. Haskins Laboratories Status Report on Speech Research SR-49.

Studdert-Kennedy, M., and Shankweiler, D. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Am. 48, 579-594.

Tobey, E. A., Cullen, J. K., Jr., and Rampp, D. L. (1976) Performance of children with auditory-processing disorders on a dichotic stop-vowel identification task. Paper presented at the Fourth Annual Meeting of the International Neuropsychological Society, Toronto, Ontario, Canada, February 1976.

Acoustic Correlates of Perceived Prominence in Unknown Utterances*

Jane H. Gaitenby and Paul Mermelstein

ABSTRACT

Intensity, fundamental frequency, and syllable duration all contribute to the perception of relative prominence among the syllables of continuous speech. These acoustic parameters were studied for their relative ability to predict syllabic prominence in a corpus of 24 imperative sentences by four talkers. The sentences were constructed with controlled syntax and limited vocabulary, as may be the case for speech communication with machines. Of the individual prosodic parameters, the best predictor of perceptual prominence was the maximum frequency-weighted intensity value for the syllable, relative to the maxima of the neighboring syllables. Duration and fundamental frequency were significantly poorer prominence predictors. A linear combination of relative intensity and duration was the best multi-parameter predictor. In polysyllabic words, perceived relative prominence ratings agreed with the intrinsic lexical stress patterns in essentially all cases. When prominence was predicted from relative intensity measurements, it agreed with the lexical stress contours for 90 percent of the words; combined relative intensity and duration brought the agreement to 92 percent.

INTRODUCTION

Prosodic features structure the speech signal at the suprasegmental level. They serve to organize sequences of syllables into words and phrases. Lexical stress is an important cue to word identity, and automatic stress indication for hypothesized syllable sequences can be expected to assist in the determination of the corresponding word sequences in speech recognition.

Duration, intensity, and the fundamental frequency contour have been previously suggested as acoustic correlates of linguistic stress (Mol and Uhlenbeck, 1956; Bolinger, 1958; Lehiste and Peterson, 1959; Lieberman, 1960). Since stress or prominence judgments can be considered to be associated with the individual syllables, it is of interest to obtain syllable-based measures for the prosodic parameters. If prominence predic-

---

*This paper has been submitted to the Journal of the Acoustical Society of America.

Acknowledgement: We appreciate the assistance of Loretta Reiss in processing the data, and the direction of the recording sessions by Lea Donald.

[HASKINS LABORATORIES: Status Report on Speech Research SR-49 (1977)]

tors can be based on single measurements per parameter rather than parameter contours, a significant data-reduction can be attained. Such measurements would truly reflect the suprasegmental aspects of the prosodic parameters if based on automatically derived syllable-sized units without regard to the segmental structure of such units. In the experiment to be described, we have characterized the duration, intensity, and fundamental frequency contours in terms of differences between adjacent syllables in the duration of the voiced subsegment of the syllable, the peak frequency-weighted intensity, and the peak fundamental frequency; and have determined the effectiveness of these measurements, individually and in combination, in predicting syllabic prominence and lexical stress for a limited amount of speech.

The relative importance of the three acoustic correlates in signaling stress in English has also received much attention. Conflicting claims abound, perhaps due to the different types of speech materials studied by the various investigators (Fry, 1958; Lieberman, 1967; Lehiste, 1970). Information concerning the phonemic content of acoustic segments is frequently signaled by a number of distinct acoustic features. Some features are necessary, others are optional. In the appropriate context, and when appropriate values are assigned to all the other features, variation of the value of each optional feature generally suffices to change the phonetic identity of the segment. It is not surprising that a similar situation is found for prosodic features. Sometimes one feature carries a heavier information load, sometimes another.

This paper is concerned with the acoustic correlates of syllable prominence (including lexical stress) in speech spoken with a limited vocabulary and with controlled syntax, as if to a speech-understanding automaton. Speech-understanding systems, for the foreseeable future, will not be able to recognize and respond to utterances selected from a natural language in its entirety. The necessity of controlling the vocabulary and syntax of the acceptable utterances will impose its own influences on the prosody of the spoken materials. The reported experiments therefore analyze the acoustic correlates of prominence in just such utterances. Generalization of the results to other modes of speech--such as fluent conversation or script readings--may not be warranted.

Background

Prosodic features have not yet been widely exploited for purposes of automatic speech recognition, although suggestions have been frequent that such features would prove useful. This lack of exploitation is due to the variable and intricate nature of the prosodic parameters in continuous speech, and to the consequent comparative rarity of publications that quantitatively describe intonation, rhythm, and rate in extensive speech samples.

Progress has been reported, however, in acoustic detection of stress and closely associated phenomena such as juncture. An outstanding example is the series of studies of American English prosodics begun by Medress, Skinner, and Anderson (1971) and continued by Lea and colleagues (1972, 1973a, 1973b, 1975, 1976a, 1976b). Lea's latest report concludes that stress is best determined by combinations of prosodic cues, of which long chunks of high

energy and local increases in fundamental frequency are the most effective. Lea has reported results ranging from 63 percent to 92 percent correct location of perceived stresses, depending on the corpus examined.

In the same study, Lea reports that the "primary simple cues to stress are (in order of increasing effectiveness): high intensity in the stressed vowel, long durations of stressed vowels or syllabic nuclei, and high F_0 [fundamental frequency] values in the stressed vowel." Lea is not explicit on the details of his stress detection method, but it appears to be based on absolute measurements. One result of the study to be described here in which relative values are used is that the reverse ranking was obtained.

Evidence was presented by Gaitenby (1974, 1975) that lexical stress in fluently-read speech is, in the majority of cases, predictable by summing weighted syllabic data for peak frequency, intensity, and duration of voicing. Although the summation method appeared to have promise for automatic stressed syllable location in words and phrases, a prerequisite to its use, as Gaitenby implied, is the creation of an algorithm for detecting syllable boundaries. Mermelstein (1975) demonstrated that automatic segmentation of syllables in continuous speech was feasible with small error rates. This suggested that automatic prominence indication might possibly be attained through assignment of an integrated prominence measure to the individual syllables.

The present experiment was undertaken to examine further the effectiveness of individual and combined prosodic parameters in locating stressed syllables. An additional consideration was an attempt to apply syllable segmentation as the first step in arriving at reliable automatic prominence detection for speech recognition purposes.

METHOD

Speech Materials

In order to record samples of speech that more closely resemble spontaneous utterances than do direct script readings, we instructed a group of talkers to create sentences for themselves, although constraints were put on the form and content of their utterances. Each talker was given a state-transition chart constraining the syntax and vocabulary of the sentences to be spoken. This diagram (shown in Figure 1) confined the syntax so that the utterances resembled commands that may be given to a computer-based robot. The vocabulary was correspondingly limited. Each talker was required to construct his or her sentences by reading left to right across the diagram, selecting words from successive columns. The talker was instructed to speak each selected sentence aloud a time or two, as rehearsal, and then to deliver the sentence for the actual recording without referring back to the diagram. We hoped that these instructions would result in some degree of spontaneity in the recorded sentences.

Four talkers were used, two men and two women, all native speakers of American English. Each talker recorded a minimum of six sentences in a single session. Twenty-four sentences were selected for analysis, four by Talker 1, six each by Talkers 2 and 3, and eight by Talker 4.

FIGURE 1

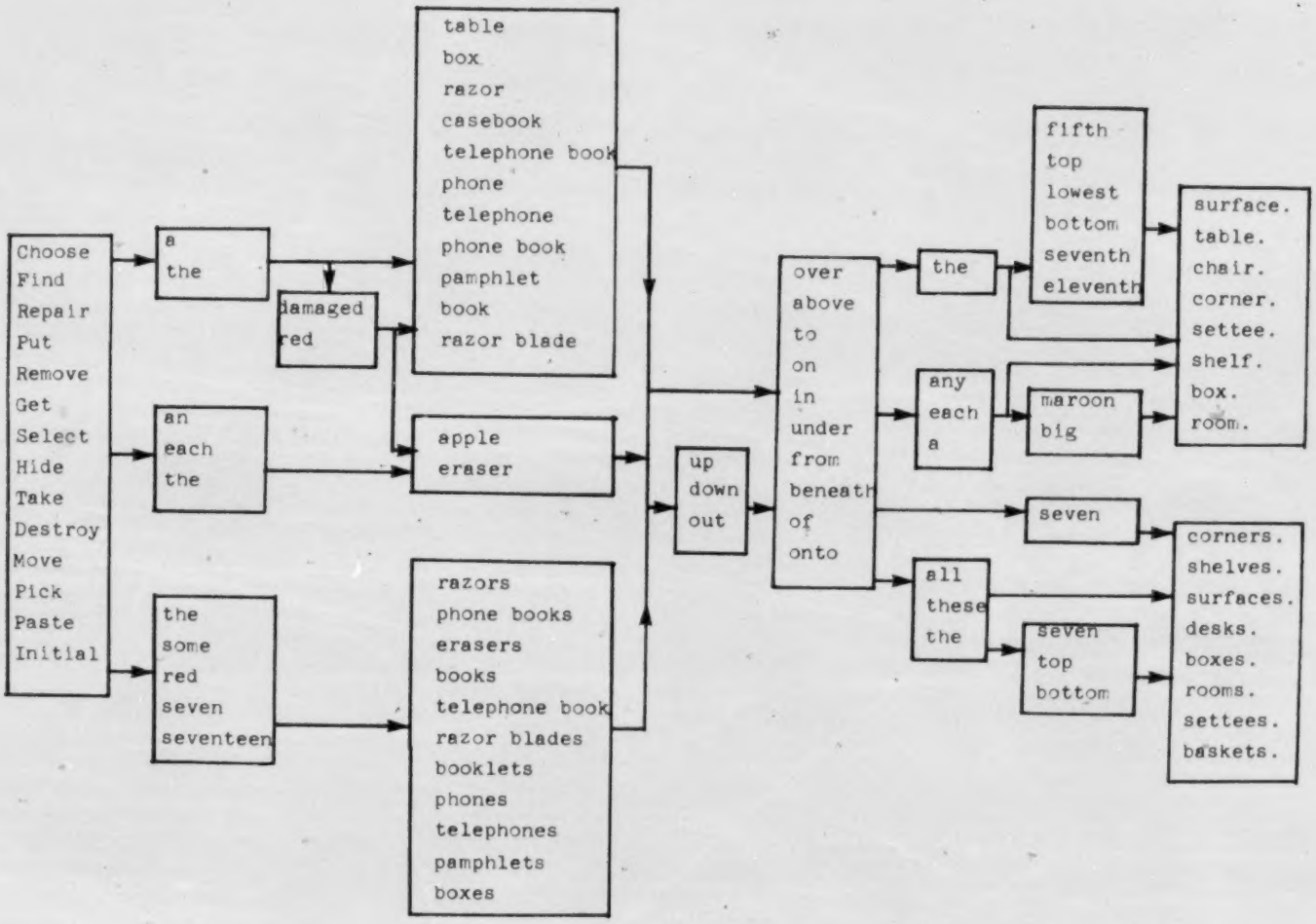


Figure 1: Syntax and vocabulary guide for sentence productions. The accompanying instructions were: 1) Assemble a sentence by reading aloud from left to right, following any connected path; 2) Rehearse the sentence aloud once or twice; 3) Say the sentence in a normal manner. (For subsequent sentences, repeat the process.)

Perceptual Measurements

Perceptual prominence judgments were collected to provide a standard to which the acoustic measurements could be compared. The taped sentences were presented to listeners in the following formal test of perceived prominence. The subjects were told to mark the more prominent syllable of each pair of adjacent syllables (syllable A vs. B, B vs. C, etc.) in every sentence. Although individual subjects were required to make binary judgments of prominence, indications of intermediate stress emerged in the pooled results. This approach to listeners' stress judgments had been useful in a test we had made in 1975 (unpublished) and a similar method was used by Lea et al (1973b) in evaluating perceived stress. (Stress and prominence may be considered equivalent terms when judgments are made between contiguous syllables of running speech.)

Five listeners were chosen from the laboratory staff. The test was taken by each subject individually, using headphones in a quiet environment, proceeding at his or her own pace. Each received a typescript of the spoken sentences on which the overlapping sequential syllable pairs were indicated, for example: [damaged, maged ta, table]. Listeners were instructed, in Listening Test 1, to write down the syllable heard as prominent in each pair. They were allowed to play the 24 taped sentences in any order, and to listen as many times as necessary to arrive at judgments. Four out of the five subjects finished the test in less than an hour, at a single sitting. None of the five found the test difficult. An additional listener had found the task impossible, and was not included in the final group.

To establish the consistency of listeners' prominence judgments, the experiment was repeated. Listening Test 2, the same as Test 1 except that the prominent syllable had to be check-marked rather than written out, was presented to each of the subjects between a week and a month after the first test. The results of Tests 1 and 2 appear as Appendices 1A and 1B, 1C and 1D.

In both tests, the initial and final syllable data in all sentences were doubled to compensate for the fact that syllables in those positions could receive only half the number of judgments received by the remaining syllables. The maximum number of votes that a syllable could receive in either test was 10, resulting from two comparisons, one with the preceding and one with the following syllable, by each of the five listeners. Ninety-five percent of the 274 syllables received the same number of pooled prominence judgments in Test 2 as in Test 1. For only four syllables did as many as three judgments (out of 10) shift in the second test. The consistency of invididual listeners in making judgments ranged from 87 percent to 91 percent, averaging 90 percent. In the pooled results of both listening tests, more than half (54 percent) of the syllables received unanimous judgments. The extent of inter-listener agreement is illustrated by the overall correlation between the most and least consistent listeners: 0.86. Judgment consistency was significantly higher for the speech of Talker 1 and lower for Talker 2 than for the two other speakers. Similar talker differences appeared in the correlations of the acoustic parameters with the perceptual data, as shown in Table 1. The strongest conflicts in judgments occurred in syllable pairs having comparable potential for prominence.

Examples are: big box, beneath all, all surfaces, in which inherently stressed syllables abut--and pamphlets in, over a, under the, that are pairs of normally unstressed syllables. The majority of these "high conflict" syllables either flanked a pause or were approximately equal in duration, or both.

TABLE 1: Correlation coefficients for acoustic parameters and perceived prominence.

Talker	1	2	3	4	Overall
No. of Sentences	4	6	6	8	24
No. of Syllables	47	69	63	95	274
Relative Intensity	0.76	0.70	0.74	0.70	0.70
Relative Duration	0.70	0.44	0.53	0.45	0.52
Relative Frequency	0.44	0.21	0.38	0.55	0.38
Relative Intensity & Relative Duration	0.81	0.73	0.77	0.77	0.77

Acoustic Measurements

The data were automatically segmented into syllable-sized units, using minima in a frequency-weighted intensity function as likely syllable boundaries (Mermelstein, 1975). The intervals within the syllabic units manifesting voicing as evidenced by a significant amount of low frequency energy (0-300 Hz) were next delimited. (It should be noted that the "syllabic unit" is not necessarily exactly equivalent to the perceived syllable where phonological and lexical criteria may play a significant role.) We attempted to weight the intensity function when integrated over frequency so that it approximated perceptual loudness. The weighting function was flat between 500 Hz and 4 kHz and dropped off at 12 dB/octave outside these frequencies. The maximum of this weighted intensity function over the voiced portion of the syllabic unit was assigned as the peak intensity of the syllable. Fundamental frequency values were computed for voiced intervals using an autocorrelation-based pitch extraction program (Lukatela, 1973), and the peak frequency for the syllable was determined. The algorithm-based measurements were cross-checked against wideband spectrograms of all 24 sentences generated with the Digital Pattern Playback (Nye, Reiss, Cooper, McGuire, Mermelstein and Montlick, 1975). The spectrograms, each hard copy displaying 1.6 sec of speech, were augmented by frequency and weighted intensity curves.

Up to this stage, data were collected without knowledge of the specific verbal content of the speech material. To correct possible syllabification errors output by the segmentation program, the recordings were then listened to. It was found that the algorithm had successfully detected 93 percent of the 274 syllables. The errors were the following: 9 cases in which one syllable was subdivided into 2, one case of 1 syllable subdivided into 3 (in [bleɪd] of "razorblade"), and seven two-syllable sequences that were not

divided. The most frequent cause of more syllables having been indicated by the program than were heard, was the presence of a clear dip in intensity within a diphthong. Too few syllables had been indicated most often in two-syllable sequences in which at least one syllable was unstressed, and in which the phone at the common boundary between the syllables was a semivowel (/r/ or /l/) or syllabic /n/. These errors were hand-corrected. In addition, syllabic units shorter than 51.2 ms were discarded, being deemed too brief to have syllable status in the particular utterances.

The acoustic measurements were converted to units similar to those used in Gaitenby's 1975 study, namely: peak fundamental frequency - 4 Hz, duration of voicing - 12.8 ms, and peak intensity - 1 dB. The parametric data, in these units, are given in Appendices 2A and 2B, 2C and 2D. [Talkers 1 and 2 (the female speakers) had fundamental frequency ranges that were much higher than those of the male talkers. To take this into consideration, the F_0 data for the lowest peak in each sentence became the baseline for the frequency measurements.]

RESULTS

Correlation coefficients between each measurement and the perceptual prominence scores were selected as indicators of the effectiveness of any one measurement in predicting perceptual prominence. (A preliminary result was that absolute intensity predicted prominence at approximately the same rate as the parameter summing method mentioned in the Background section of this report. The correlation coefficient for absolute intensity and perceived prominence was 0.54.) Since the perceptual judgments were determined relative to the prominence of the neighboring syllables, the intensity, frequency and duration measurements were converted to relative measures through the use of the following local difference function on groups of three consecutive syllables:

$$M_x = 2M_x - (M_{x-1} + M_{x+1})$$

where M_x' is a relative measure for syllable x (peak frequency, peak intensity, or duration), and M_x is the absolute measure for the same variable. For the initial and final syllable we used

$$M_x = 2(M_x - M_{x+1})$$

and

$$M_x = 2(M_x - M_{x-1})$$ respectively.

The resulting correlation coefficients were given in Table 1. It is apparent that of the individual measurements, relative intensity is the single best correlate of relative prominence. In terms of the overall results, duration and fundamental frequency are the worst correlates, in that order. However, when we look at the correlation coefficients for the texts of individual talkers, duration is less highly correlated with prominence than frequency for one talker. There appear to be significant differences in the way the various talkers encode the prominence information in terms of the three prosodic parameters.

We next investigated whether a linear combination of the two most effective parameters would prove more useful than either of them alone. The two best individual parameters, relative intensity and duration, showed a correlation coefficient of 0.3 with respect to each other. Based on multiple regression techniques (McNemar, 1969), the resulting best estimator for prominence from relative intensity and duration was determined to be

$$P_{est} = 0.59\ I_{rel} + 0.32\ D_{rel}$$

where I_{rel} and D_{rel} are the relative intensity and duration measurements, and P_{est} is the estimated relative prominence. The correlation of this new estimate with the prominence judgments was 0.77.

To judge the effectiveness of the above correlation figure, we attempted to determine the disagreement to be expected between the prominence ratings of different listeners. It is unlikely that the agreement between the overall prominence judgments and that predicted from acoustic measurements can exceed the agreement between prominence judgments of individual listeners. Since each listener judged the spoken data twice, consistency measures were available on the judgments by each listener. The most consistent and least consistent listeners were selected to illustrate the range of judgments one can expect. The overall correlation between the judgments of these two subjects was 0.86. This figure, then, represents a rough upper limit to which the correlation between the best estimate of prominence and its judgment may be compared. Evidently, relative intensity and duration are quite effective when used in combination to predict relative prominence. Relative intensity alone is slightly poorer.

DISCUSSION

The results of the correlation analyses show that the ranking of the single parameters as prominence cues is intensity first, duration second, and fundamental frequency third--the reverse order of that found by Fry (1958) and Lea (1976a). A main difference between this analysis and those, aside from the type of intensity measurement, is that the present data are values relative to the adjacent syllables. Another difference is that many of our sample sentences were delivered rather slowly and hesitantly. The sizable pitch excursions and strongly contrasting durations that may sometimes accompany fluent speech tend to disappear in utterances that are hesitantly or cautiously produced, with rhythmic phrasing lessening as stress tends to be applied more evenly to all of the words in slow speech. When frequency and duration are "under-used" as cues to prominence, the reliability of intensity as a stress signal may increase due to the well-known trading effects among the parameters.

Perceived relative prominence ratings agreed with intrinsic (that is, dictionary) lexical stress in essentially every case. For these multisyllabic words we attempted to predict lexical stress from acoustic measurements. Relative intensity correctly indicated lexical patterns in 90 percent of the (multisyllabic) words. Two words were responsible for 5 of the 8 errors found: "beneath" and "maroon." Since /i/ or /u/ appeared in the stressable syllable of these two words, normalization with respect to the average peak intensity found for these vowels was tried, but without significant result.

It was noted that error words "beneath" and "maroon" usually preceded a pause and were accompanied by a fall in peak F_0 and a large rise in duration relative to the preceding syllable. Duration was plainly the solitary suprasegmental feature signaling prominence in those cases. Also noted, in passing, was the fact that both of these words and "settee," another word producing an error, are intrinsically stressed on the final syllable.

When the linear combination of relative intensity and relative duration was used as a predictor of major lexical stress, the number of errors fell to 6; and in predicting secondary stress, one error occurred in the word "razorblade." In total, the speech sample contained 77 polysyllabic words (37 different words, including 5 in both singular and plural forms). There were 11 trisyllabic tokens and 66 disyllables. Using the intensity and duration combination for stress prediction, the words with errors were, as before, all disyllabic: "beneath" (twice), "under" (twice), "any" and "settee." Again the errors occurred often in prepausal words containing the phone /i/ in the normally stressed syllable. Four out of 6 errors were in function words. All error words occurred late in the sentences. It was noted too that the meaning of both "beneath" and "under" is low, a factor that might influence their prosodics to some extent. Combined intensity and duration predicted the lexical stress patterns in 97 percent of the polysyllabic content words and 69 percent of the function words. For all polysyllables, the intensity-duration combination's stress prediction rate was 92 percent. This figure equals the highest prediction rate for perceived stresses achieved by Lea (1976a) and by Sargent (1975). Our 2 percent gain in overall stress prediction, achieved by the inclusion of duration as well as intensity data, may be hardly worth the increased complexity of the algorithm. Alone, as has been shown, frequency-weighted intensity is a highly reliable stress predictor.

Lea reported elsewhere (1976b pp. 6-8) that his approach had succeeded in detecting syllables at an 81 percent rate in a corpus that consisted of 15 statements, questions, and commands, and that 63 percent of the stressed syllables had been located correctly. The 24 sentences we have examined here are comparable in length to those used by Lea, but represent only commands, and might be considered more simple in syntax. A precise comparison of results is therefore difficult. Nevertheless, our overall syllable detection rate was 93 percent, and the correlation of 0.77 between relative prominence predicted via intensity and duration, and perceptually judged, suggests that 85 to 90 percent of all syllables judged prominent would be located. Automatic stress assignment requires the construction of a decision rule based on acoustic measurements such as those used here. In polysyllabic words the simplest rule is to assign major lexical stress to the syllable in the word found most prominent. For monosyllabic words, a simple threshold on the relative prominence may suffice for stress assignment. Prominence measures can, additionally, serve to predict the clarity with which the acoustic information can be expected to be manifested.

A few peripheral observations about the sample sentences are worthy of mention. First, the limited syntatic structure used in our sample sentences was meant not only to resemble commands to a robot, but also to reveal structural relationships with prosodic features. So far, aside from certain long pauses, evidence of regular prosodic reflections of syntax has not been

found in the data. This result may not be surprising in view of the fact that only syllable peak data were examined and the utterances were predominantly slow.

Second, pause length was extremely variable, both within and across talkers, ranging up to 1.8 secs. Seventeen of the 24 sentences contained at least one pause, and Talker 3 produced five of the sentences lacking any pause. If more than one pause occurred in a sentence, the first was the longest. Most of the pauses took place after the first noun (which preceded the adverbial phrase) and thus appeared to have syntactic relevance. No pauses occurred at any earlier location in a sentence. There were a few cases of pause introduced between an adjective (or article) and the final noun. In this position the word "a" was pronounced [eɪ] and "the" became an elongated [ðʌ] or [ði]. Such hesitation effects had several possible causes: the spatial design of the diagram given to the talkers as a guide to their productions, the restricted vocabulary, and the constraints of the speaking task as a whole. Average time intervals between stressed syllables and pause length showed no dependable relationships.

Finally, Table 2 shows the intensity and frequency ranges for the four individual talkers. The voices of the female speakers were "typically" high; the men's were low. The women displayed not only a larger frequency range, but also a smaller intensity range than that of both men. As expected, the range and ratios of voicing duration were similar for all four talkers. Generally speaking, there was as much variation in the absolute duration for a given word within a single talker's speech as there was across the talkers.

TABLE 2: Ranges of intensity and frequency, by Talker

	Talker	Intensity	Frequency
#1	Female	11.2 dB	108 Hz
#2	"	12.8	110
#3	Male	14.7	72
#4	"	16.3	83

CONCLUSIONS

Relative prominence of syllables in continuous speech may be predicted from syllable-based measurements with a reliability approaching the agreement between individual listeners. The most and least consistent listeners in the perceptual test of prominence showed a mutual correlation of 0.86. This figure is a standard against which the acoustic predictions of prominence may be evaluated. Of the three individual prosodic parameters, a relative measure of spectrally-weighted intensity correlates most highly (0.70 overall) with perceived prominence in the (mostly) slow speech sample. Syllable prominence is predicted more closely, however, by a combination of relative intensity and relative duration of voicing, with an overall correlation of

0.77. This combination predicts lexical stress in 92 percent of polysyllabic content and function words.

The sample we have discussed includes utterances by only four talkers, two men and two women; therefore the observation made on male versus female prosodic differences can be considered only suggestive. The implication from our very limited data is that female speakers have both a wider frequency range and a narrower intensity range than males. Further research is plainly needed on prosodic differences in male versus female speech. One question is: to what extent are these differences affected by socio-linguistic factors?

Research is also needed on the extent to which the use of the separate prosodic cues to prominence change with increasingly rapid speech for a variety of speakers. A persisting related question is how speech material and other factors influence speech rate and segmental duration. The dependence of vowel duration on speech mode, for example, is highlighted in Harris and Umeda (1974) where it is concluded that the role of prosody seems to be very different in carrier phrases as opposed to connected text.

The present results provide further quantitative evidence that different talkers may use their prosodic resources for prominence in different ways, some using more intensity or frequency variations, others, more durational cues. Lieberman and Michaels (1962) have made a similar observation, that individual talkers show prosodic differences in expressing emotional attitudes. Nevertheless, in the present speech sample, the pattern of relative intensity is the single feature shared dependably by all four talkers in signaling prominence. Carefully spoken sentences--like those examined in this report--may be the most recognizable form of connected speech input to computers for some time. We suggest the use of the simple frequency-weighted intensity measure for prominence prediction in this type of man-machine communication task.

T1S1 Select each eraser down beneath all these seven desks.

Test 1	0	6	9	0	10	0	10	0	6	9	2	8	0	5	
" 2	0	5	10	0	10	0	10	0	7	8	1	9	0	5	25
Tot.	0	11	19	0	20	0	20	0	13	17	3	17	0	~~10~~	

T1S2 Repair the damaged pamphlet above the eleventh chair.

1	0	10	0	10	0	10	5	0	10	4	1	10	0	5	
2	0	10	0	10	0	10	5	0	10	4	1	10	0	5	20
Tot.	0	20	0	20	0	20	10	0	20	8	2	20	0	~~10~~	

T1S3 Take the apple up from a big box.

1	5	0	10	0	9	6	0	9	1
2	5	0	10	0	9	5	1	7	3
Tot.	20 ~~10~~	0	20	0	18	11	1	16	8 ~~4~~

T1S4 Move the phonebooks out beneath seven shelves.

	5	0	10	0	10	0	8	7	0	5
	5	0	10	0	10	0	7	8	0	5
Tot.	20 ~~10~~	0	20	0	20	0	15	15	0	20 ~~10~~

T2S1 Find a damaged razorblade down beneath all seven desks.

	5	0	10	0	10	0	5	10	0	5	9	6	0	5
	5	0	10	0	10	0	5	10	0	7	7	6	0	5
Tot.	20 ~~10~~	0	20	0	20	0	10	20	0	12	16	12	0	20 ~~10~~

T2S2 Remove each eraser up from any maroon box.

	0	8	7	0	10	0	9	2	3	4	1	6	4
	0	7	8	0	10	0	10	0	10	4	1	6	4
Tot.	0	15	15	0	20	0	19	2	13	8	2	12	16 ~~8~~

T2S3 Paste some booklets up under each maroon shelf.

	5	0	10	0	8	7	0	10	0	6	4
	5	0	10	1	8	6	0	10	0	6	4
Tot.	20 ~~10~~	0	20	1	16	13	0	20	0	12	16 ~~8~~

Appendix 1A

T2S4 Hide the apple to the lowest surface.

Test 1	5	0	10	1	7	2	10	0	10	0
" 2	5	0	10	0	10	0	9	1	10	0
Tot.	20 ~~10~~	0	20	1	17	2	19	1	20	0

T2S5 Pick each casebook down under each big box.

1	3	2	10	0	9	6	0	7	5	3
2	5	1	9	0	9	6	0	8	2	5
Tot.	16 ~~8~~	3	19	0	18	12	0	15	7	16 ~~8~~

T2S6 Select some erasers down from a big shelf.

	0	9	6	0	10	0	10	4	1	7	3
	0	9	5	1	10	0	10	5	2	6	4
Tot.	0	18	11	1	20	0	20	7	3	13	14 ~~7~~

T3S1 Repair a damaged razorblade in the seventh box.

	0	10	0	10	0	10	0	10	5	0	10	1	4
	0	10	0	10	0	10	0	9	5	1	10	1	4
Tot.	0	20	0	20	0	20	0	19	10	1	20	2	16 ~~8~~

T3S2 Select an apple from all seven shelves.

	0	10	0	10	1	4	9	6	0	5
	0	10	0	10	1	4	9	6	0	5
Tot.	0	20	0	20	2	8	18	12	0	20 ~~10~~

T3S3 Paste some books down onto all surfaces.

	5	0	8	7	0	8	7	0	5		
	5	0	8	6	6	0	6	9	1	4	
Tot.	20 ~~10~~	0	16	13	11	0	14	16	1	18 ~~9~~	

T3S4 Hide a razor in the top room.

	5	0	10	3	7	0	7	3
	5	0	10	2	7	1	7	3
Tot.	20 ~~10~~	0	20	5	14	1	14	12 ~~6~~

Appendix 1B

T3S5 Get the apple down from the top corner.

Test 1	5	0	10	0	10	4	1	6	9	0	
" 2	5	0	10	0	10	5	0	5	10	0	
Tot.	~~10~~ 20	0	20	0	20	9	1	11	19	0	

T3S6 Pick seven telephones up from any big room.

	1	9	0	10	0	6	9	2	8	1	4	5
	2	8	0	10	0	6	9	1	9	0	5	5
Tot.	~~3~~ 6	17	0	20	0	12	18	3	17	1	9	~~10~~ 20

T4S1 Paste each damaged casebook up over a shelf.

	0	9	6	0	10	0	9	6	4	1	5
	2	7	6	0	10	0	9	6	1	4	5
Tot.	~~2~~ 4	16	12	0	20	0	18	12	5	5	~~10~~ 20

T4S2 Get seven erasers out beneath the settee.

	1	9	1	4	10	0	10	0	10	0	6	4
	0	10	2	3	10	0	10	0	10	0	5	5
Tot.	~~1~~ 2	19	3	7	20	0	20	0	20	0	11	~~9~~ 18

T4S3 Remove each damaged apple out under the bottom table.

	0	5	7	8	0	10	0	10	5	0	5	9	1	10	0
	0	5	7	8	0	10	0	9	5	1	5	10	0	10	0
Tot.	0	10	14	16	0	20	0	19	10	1	10	19	1	20	0

T4S4 Destroy red pamphlets in the seven books.

	0	7	6	7	4	5	1	10	0	5
	0	6	6	8	2	7	1	10	0	5
Tot.	0	13	12	15	6	12	2	20	0	~~10~~ 20

T4S5 Hide the telephone books up under the corner.

	5	0	10	0	6	5	9	5	1	4	10	--
	5	0	10	0	6	4	10	5	2	3	10	--
Tot.	~~10~~ 20	0	20	0	12	9	19	10	3	7	20	(Syll. lost)

Appendix 1C

T4S6 Select each damaged razorblade in every surface.

Test 1	0	5	9	7	0	9	0	10	1	9	0	10	0	
" 2	0	5	9	6	0	10	0	10	0	10	0	10	0	
Tot.	0	10	18	13	0	19	0	20	1	19	0	20	0	

T4S7 Put the damaged table up onto the surface.

	5	0	10	0	10	0	8	7	1	4	10	0	
	5	0	10	0	10	1	8	6	1	4	10	0	
Tot.	~~10~~ 20	0	20	0	20	1	16	13	2	8	20	0	

T4S8 Hide some red apples up under the baskets.

	5	0	7	8	0	9	6	0	5	10	0
	5	0	6	9	0	9	6	2	3	10	0
Tot.	~~10~~ 20	0	13	17	0	18	12	2	8	20	0

Appendix 1D

T1S1	Select each eraser down beneath all these seven desks.													
F_0	24	20	24	24	17	13	15	20	14	15	12	23	12	18
Dur.	7	13	17	10	16	20	28	11	22	22	16	10	16	15
Int.	11	12	4	3	9	1	7	5	3	8	1	5	0	4
T1S2	Repair the damaged pamphlet above the eleventh chair.													
F_0	25	25	21	18	18	22	14	14	13	17	4	11	11	19
Dur.	12	34	14	15	17	18	12	10	24	26	6	26	11	20
Int.	6	12	9	3	6	8	6	4	10	5	0	7	0	3
T1S3	Take the apple up from a big box.													
F	29	32	15	14	16	10	15	19	15					
D	17	9	15	21	12	32	19	21	15					
I	12	7	5	0	6	4	1	7	2					
T1S4	Move the phonebooks out beneath seven shelves.													
F	27	22	23	10	23	20	13	21	13	20				
D	29	7	24	14	21	6	21	9	12	26				
I	8	6	8	8	9	5	3	7	0	6				
T2S1	Find a damaged razorblade down beneath all seven desks.													
F	22	23	15	12	16	11	9	10	10	8	8	10	6	13
D	18	11	14	14	20	10	22	24	8	18	19	12	13	16
I	10	4	8	1	9	0	2	12	5	0	6	11	2	11
T2S2	Remove each eraser up from any maroon box.													
F	15	21	15	24	13	12	12	17	6	8	11	12	7	
D	10	21	12	5	17	15	11	21	10	21	9	26	14	
I	3	7	4	3	12	2	11	5	5	2	0	1	8	
T2S3	Paste some booklets up under each maroon shelf.													
F	21	21	16	3	9	9	7	9	11	7	10			
D	19	14	9	12	9	15	10	13	12	31	16			
I	13	2	8	7	10	10	3	0	4	3	10			

Appendix 2A

T2S4	Hide the apple to the lowest surface.													
F	18	11	14	14	11	5	7	7	7	25				
D	21	10	13	17	13	10	18	8	10	4				
I	17	11	13	4	8	13	14	13	7	0				
T2S5	Pick each casebook down under each big box.													
F	46	42	41	24	31	30	28	3	31	2				
D	7	10	12	12	22	14	10	14	18	17				
I	8	6	10	12	12	5	5	0	6	9				
T2S6	Select some erasers down from a big shelf.													
F	17	31	34	23	25	13	26	29	16	21	25			
D	4	14	5	12	20	13	33	10	18	24	18			
I	0	10	5	4	9	0	11	4	2	3	8			
T3S1	Repair a damaged razorblade in the seventh box.													
F	12	17	12	13	12	13	11	6	7	7	13	12	7	
D	11	23	9	20	8	18	11	24	12	12	11	11	21	
I	9	13	11	12	5	12	4	7	4	6	11	7	11	
T3S2	Select an apple from all seven shelves.													
F	9	14	21	23	14	7	12	12	10	11				
D	6	12	9	17	9	14	25	11	12	21				
I	9	16	11	14	8	10	12	12	0	12				
T3S3	Paste some books down onto all surfaces.													
F	15	18	17	15	8	13	8	13	9	7				
D	14	19	17	24	17	10	11	9	(3)	6				
I	9	1	6	4	3	0	5	0	0	0				
T3S4	Hide a razor in the top room.													
F	18	5	13	5	5	5	11	9						
D	18	11	24	10	16	7	17	24						
I	18	9	13	6	6	0	15	2						

Appendix 2B

T3S5	Get the apple down from the top corner.																
F	14	21	9	8	11	10	4	6	13	4							
D	13	9	19	16	29	11	7	16	11	8							
I	15	8	15	8	16	5	0	15	11	4							
T3S6	Pick seven telephones up from any big room.																
F	14	15	10	16	7	8	11	10	8	15	8	10					
D	9	12	14	12	9	29	15	18	11	13	20	34					
I	15	16	9	15	15	10	15	9	5	0	4	1					
T4S1	Paste each damaged casebook up over a shelf.																
F	22	30	16	17	10	0	13	11	9	7	1						
D	14	14	16	16	9	10	9	17	7	7	11						
I	9	6	16	8	9	0	12	5	2	2	7						
T4S2	Get seven erasers out beneath the settee.																
F	13	19	22	16	10	3	19	8	9	0	6	9					
D	11	10	10	12	16	12	19	12	12	5	8	15					
I	12	18	12	13	13	3	16	0	4	1	11	2					
T4S3	Remove each damaged apple out under the bottom table.																
F	12	22	21	14	13	13	1	12	11	11	5	9	13	6	0		
D	13	21	15	19	7	14	14	17	12	13	18	13	14	16	10		
I	3	11	11	16	7	14	4	17	8	4	0	14	10	5	2		
T4S4	Destroy red pamphlets in the seven books.																
F	8	17	11	13	13	4	4	6	6	7							
D	7	27	21	19	11	13	8	12	27	11							
I	0	14	14	9	3	1	0	11	0	5							
T4S5	Hide the telephone books up under the corner.																(Syll. lost)
F	14	11	17	13	4	4	9	5	6	3	13						
D	15	5	10	9	23	12	13	24	7	4	10						--
I	20	1	21	21	13	11	20	13	5	0	14						--

Appendix 2C

T4S6	Select each damaged razorblade in every surface.																(devcd.)
F	9	13	23	9	10	0	4	4	6	5	4	12					
D	5	14	13	15	12	10	15	23	6	14	12	10	8				
I	5	17	7	14	5	9	0	5	0	8	2	9	0				
T4S7	Put the damaged table up onto the surface.																
F	23	14	17	13	22	9	10	12	8	5	13	0					
D	8	11	13	9	17	14	8	13	8	7	11	8					
I	11	2	14	5	8	4	12	6	1	0	11	0					
T4S8	Hide some red apples up under the baskets.																
F	16	13	11	11	2	9	10	8	0	8	0						
D	16	14	16	16	15	12	14	8	9	16	6						
I	14	8	9	12	0	7	4	0	0	8	0						

Appendix 2D

REFERENCES

Bolinger, D. (1958) A theory of pitch accent in English. Word 14, 109-149.

Fry, D. B. (1958) Experiments in the perception of stress. Lang. Speech 1, 126-152.

Gaitenby, J. H. (1974) The elastic syllable: An acoustic view of the stress-intonation link. J. Acoust. Soc. Am. 56, S32 (A).

Gaitenby, J. H. (1975) Stress and the elastic syllable: An acoustic method for delineating lexical stress patterns in connected speech. Haskins Laboratories Status Report on Speech Research SR-41, 137-152.

Harris, M. S., and Umeda, N. (1974) Effect of speaking mode on temporal factors in speech: Vowel duration. J. Acoust. Soc. Am. 56, 1016-1018.

Lea, W. A., Medress, M. F., and Skinner, T. E. (1972) Basic algorithms and stress studies. Sperry Univac Rep. PX 7940.

Lea, W. A., Medress, M. F., and Skinner, T. E. (1973a) Syntactic segmentation and stressed syllable location. Sperry Univac Rep. PX 10232.

Lea, W. A., Medress, M. F., and Skinner, T. E. (1973b) Relationships between stress and phonemic recognition results. Sperry Univac Rep. PX 10430, p. 17.

Lea, W. A., Medress, M. F., and Skinner, T. E. (1975) A prosodically guided speech understanding strategy. IEEE Trans. Acoust., Speech, and Signal Proc. ASSP-23, 30-38.

Lea, W. A. (1976a) Acoustic correlates of stress and juncture. Sperry Univac Rep. PX 11693.

Lea, W. A. (1976b) Listeners' perceptions of selected English stress patterns. Sperry Univac Rep. PX 11711.

Lehiste, I., and Peterson, G. E. (1959) Vowel amplitude and phonemic stress in American English. J. Acoust. Soc. Am. 32, 428-435.

Lehiste, I. (1970) Suprasegmentals. (Cambridge: MIT Press).

Lieberman, P. (1960) Some acoustic correlates of word stress in English. J. Acoust. Soc. Am. 32, 451-453.

Lieberman, P. (1967) Intonation, Perception and Language. (Cambridge: MIT Press).

Lieberman, P., and Michaels, S. B. (1962) Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J. Acoust. Soc. Am. 34, 922-927.

Lukatela, G. (1973) Pitch determination by adaptive autocorrelation method. Haskins Laboratories Status Report on Speech 33, 185-193.

McNemar, Q. (1969) Psychological Statistics, 4th edition. (New York: John Wiley).

Medress, M., Skinner, T. E., and Anderson, E. D. (1971) Acoustic correlates of word stress. J. Acoust. Soc. Am. 51, 101 (A).

Mermelstein, P. (1975) Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58, 880-883.

Mol, H. G., and Uhlenbeck, G. M. (1956) The linguistic relevance of intensity in stress. Lingua 5, 205-213.

Nye, P. W., Reiss, L. J., Cooper, F. S., McGuire, R. M., Mermelstein, P., and Montlick, T. (1975) A digital pattern playback for the analysis and manipulation of speech signals. Haskins Laboratories Status Report on Speech Research 44, 95-107.

Sargent, D. C. and K. S. Fu. (1975) Computer algorithms for the extraction and application of stress contours from continuous speech sentences. (School of Electrical Engineering, Purdue University, West Lafayette, Indiana), Report No. TR-EE 75-44.

PUBLICATIONS AND REPORTS

Cooper, Franklin S., P. Mermelstein and P. W. Nye. (1976) Speech synthesis as a tool for the study of speech production. Dynamic Aspects of Speech Production: Current Results, Emerging Problems, and New Instrumentation, U.S.-Japan Joint Seminar, Hill Top Hotel, Tokyo, Japan, 7-10 December 1976, 141-149.

Cutting, James E. and M. F. Dorman. (1976) Discrimination of intensity differences carried on formant transitions varying in extent and duration. Percept. Psychophys., vol. 20(2), 101-107.

Gay, Thomas. (1976) Cinefluorographic and electromyographic studies of articulatory organization. Dynamic Aspects of Speech Production: Current Results, Emerging Problems, and New Instrumentation. U.S.-Japan Joint Seminar, Hill Top Hotel, Tokyo, Japan, 7-10 December, 59-70.

Harris, K. S. (1976) The study of articulatory organization: Some negative progress. Dynamic Aspects of Speech Production: Current Results, Emerging Problems, and New Instrumentation. U.S.-Japan Joint Seminar, Hill Top Hotel, Tokyo, Japan, 7-10 December, 53-58.

Healy, Alice F. and M. Kubovy. (1977) A comparison of recognition memory to numerical decision: How prior probabilities affect cutoff location. Memory and Cognition, vol. 5(1), 3-9.

Mermelstein, Paul. (1977) Distance functions for speech recognition--psychological and instrumental. In Pattern Recognition and Artificial Intelligence, ed. by C. H. Chen. (New York: Academic Press), pp. 374-388.

Mermelstein, Paul. (1977) On detecting nasals in continuous speech. J. Acoust. Soc. Am. 61, 581-587.

Nye, Patrick W. (1976) Reading devices for blind people. Med. Prog. through Technolog., vol. 4, 11-25.

Repp, Bruno. (1977) Dichotic competition of speech sounds: The role of acoustic stimulus structure. J. Exp. Psychol. (HPP), 3, 37-50.

Strange, Winifred, Verbrugge, R., Shankweiler, D. P. and Edman, T. R. (1976) Consonant environment specifies vowel identity. J. Acoust. Soc. Am. 60, no. 1, 213-224.

Summerfield, Quentin and M. Haggard. (1977) Perceptual calibration for parameters of speaker differences--Measures from sequential reaction time increment studies. In Attention and Performance VI, ed. by S. Dornic. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.), pp. 261-282.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-48:*

Status Report		DDC*	ERIC*
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273
SR-45/46	January - June 1976	AD A026196	ED-123-678
SR-47	July - September 1976	AD A031789	ED-128-870
SR-48	October - December 1976	AD A036735	ED-135-028
SR-49	January - March 1977	**	**

*See next page for ordering information.

**DDC and/or ERIC order numbers not yet assigned.

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	Unclassified
	2b. GROUP N/A

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 49, January - March 1977

4. DESCRIPTIVE NOTES (*Type of report and, inclusive dates*)

Interim Scientific Report

5. AUTHOR(S) (*First name, middle initial, last name*)

Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1977	219	241

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DE-01774 BNS76-82023 HD-01994 MCS76-81034 V101(134)P-342 MDA 904-77-C-0157 N01-HD-1-2420 RR-5596	SR-49 (1977)
	9b. OTHER REPORT NO(S) (*Any other numbers that may be assigned this report*) None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT

This report (1 January - 31 March 1977) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation of its investigation, and practical applications. Manuscripts cover the following topics:

-Dissociation Spectral, Temporal Cues to Voicing Distinction in Initial Stop Consonants
-Perceptual Integration and Selective Attention in Speech Perception: Further Experiments on Intervocalic Stop Consonants
-Phonetic Recoding and Reading Difficulty in Beginning Readers
-Interactive Experiments with Digital Pattern Playback
-Function of Strap Muscles in Speech: Pitch Lowering or Jaw Opening?
-Geniohyoid and Role of Strap Muscles in Pitch Control
-Syllable Synthesis
-Articulatory Movements in VCV Sequences
-Measuring Laterality Effects in Dichotic Listening
-Simple Model of Response Selection in Dichotic Two-Response Paradigm
-Acoustic Correlates of Perceived Prominence in Unknown Utterances

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED

Security Classification

A-31408

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Cues - Spectral, Temporal, Stop Consonants Integration and Attention: Speech Perception Reading - Phonetic, Beginning Readers Spectrogram Reading, Playback Strap Muscles - Speech, Pitch, Jaw Geniohyoid, Strap Muscles, Pitch Syllable Synthesis Articulation - VCV Sequences Dichotic Listening, Laterality Response Selection - Model, Dichotic Perception, Speech, Prominence						